UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO FIN DE GRADO

Modelado de la difusión de información en Twitter

Blanca Miranda Mayans
Tutor: Javier Sanz-Cruzado Puig
Ponente: Pablo Castells Azpilicueta

JUNIO 2018

Modelado de la difusión de información en Twitter

AUTOR: Blanca Miranda Mayans
TUTOR: Javier Sanz-Cruzado Puig
PONENTE: Pablo Castells Azpilicueta

Grupo de Recuperación de Información
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2018

Resumen

La llegada de las redes sociales en línea ha traído consigo una revolución en la difusión de información. Comparado con los medios de comunicación tradicionales centralizados en unas pocas manos, con unos grandes canales radiales de transmisión, las redes se caracterizan por la capacidad de expandir información en oleadas a partir de un pequeño origen en cualquier punto de la red, llegando a un gran número de receptores en muy poco tiempo. Este fenómeno se ha convertido en objeto de estudio por su gran impacto en la sociedad.

En este Trabajo de Fin de Grado se investiga, en el campo de difusión de información en redes sociales, la estructura de cascada de los posibles caminos que ha seguido la información en su propagación por la red. El trabajo experimental del TFG se basa en datos de la plataforma social Twitter, cuyo objetivo es la publicación de tweets, pequeñas piezas de información, que otros usuarios pueden recibir y retransmitir a otros; un formato que resulta muy adecuado en este estudio.

En una primera fase se realiza una captura de datos de Twitter y se recrean los caminos posibles seguidos por la información en forma de cascadas para, posteriormente, ser analizadas con métricas de caracterización de grafos. En este análisis se estudian las distribuciones de los valores aportados por las métricas para obtener información de la topología y propiedades de estos caminos seguidos por la información.

En una segunda fase del trabajo, se estudiarán diferentes modelos de influencia, cuyo objetivo es refinar las cascadas de información escogiendo el camino más plausible de entre todos los posibles, según las hipótesis de cada modelo. Cada modelo representa la influencia que ejercen unos usuarios sobre otros en este proceso de propagación basándose en distintos criterios. En este trabajo se intenta concluir cuál de ellos se aproxima más a lo que ocurre en la realidad. Para ello, se empleará un método de simulación de propagación de tweets, que permite realizar comparaciones y análisis para obtener datos que lleven al objetivo de esta fase.

Palabras clave: red social, difusión de información, propagación, cascada, modelo de influencia.

Abstract

The arrival of social networks online has caused a revolution in the information diffusion. Compared to traditional media centralized in a few individuals, with large radial transmission channels, the networks are characterized by their ability to expand information in waves from any point of the network, reaching many receivers in a very short time. This phenomenon has become an object of study due to its great impact on society.

In this Bachelor Thesis, the cascade structure of the possible paths that information has followed in its propagation through the network is investigated, in the field of information diffusion in social networks. The experimental work in this Thesis is based on data from the social platform Twitter, whose objective is the publication of tweets, small pieces of information, that other users can receive and retransmit to others; a format that is very appropriate in this research.

In a first phase, a data capture of Twitter is made, and the possible paths followed by the information in the form of cascades are recreated and subsequently analyzed with graph characterization metrics. In this analysis, the distributions of the values contributed by the metrics are studied to obtain information of the topology and properties of these paths followed by the information.

In a second phase of the work, different influence models will be studied, whose objective is to refine the information cascades by choosing the most plausible path among all possible ones, according to the hypothesis of each model. Each model represents the influence exerted by users on others in this propagation process based on different criteria. In this Thesis we try to conclude which of them is closer to what happens in reality. For this purpose, a simulation method of tweets propagation will be used, which allows to make comparisons and analysis to obtain data that lead to the objective of this phase.

Keywords: social network, information diffusion, spread, information cascade, influence model.

Agradecimientos

En primer lugar, agradezco toda la ayuda a mis tutores Pablo y Javi.

Gracias a mi familia, especialmente a mi madre por el gran apoyo en esta última fase que tanto esfuerzo me ha costado.

Gracias a mis amigas que siempre están ahí, sobre todo a Ale, que día a día me da fuerzas.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

La difusión de información es un amplio campo de estudio en el ámbito de las redes sociales. Desde hace años, en diversos estudios se ha investigado sobre la posibilidad de crear un modelo que simule el comportamiento de la difusión de información en redes sociales.

La propagación de información a través de una estructura de grafo presenta peculiaridades específicas en comparación con otras formas de difusión, como los *mass media*. En las redes sociales en línea, las particularidades son aún más acentuadas comparadas con la comunicación presencial, o incluso la comunicación con soporte tecnológico más tradicional, como el teléfono o el correo electrónico.

La propagación de información en redes online se caracteriza por una altísima velocidad potencial, con elementos exponenciales, amortiguados por factores de decaimiento, en los que la información se propaga en oleadas, comparables a la expansión de un contagio. Tal y como muestra Newman en su trabajo (2010), la difusión de la información podría asemejarse a la propagación de enfermedades en una población, para lo cual existen modelos basados en ecuaciones diferenciales que representan los sucesos de la realidad con bastante precisión.

El estudio y la comprensión del comportamiento de estas oleadas tiene una gran trascendencia, debido al impacto que tienen en ámbitos tales como el conocimiento colectivo o la opinión pública. Por su novedad y relevancia, y algunos aspectos en cierto grado intrigantes, estos fenómenos se han convertido en objeto de amplia investigación en años recientes, impulsada por el auge de los medios sociales online a escala mundial.

La investigación en este ámbito se ha enfocado a aspectos tales como la velocidad, el impacto o el alcance de la propagación de información; y a tratar de explicar el salto desde las acciones entre usuarios a nivel micro, hasta los fenómenos a los que éstas dan lugar a nivel macro, observados con perspectiva global.

En este contexto, el trabajo que aquí se presenta se centra en la reconstrucción de los caminos seguidos por la información a partir de una observación parcial e incompleta de la comunicación entre usuarios de una red social. Sobre esta base el trabajo se enfoca, en primer lugar, en el análisis de las propiedades estructurales que describe la información a su paso, a escala media y macro. Por otra parte, se buscará explicar las dinámicas observadas mediante el postulado de modelos de interacción e influencia entre usuarios. Se comparará y valorará la verosimilitud de los modelos mediante simulaciones y su contraste con las observaciones parciales disponibles.

El estudio va a centrarse en la red social Twitter. Ésta se basa en la publicación de tweets por parte de usuarios, los cuales son retuiteados (compartir el tweet original con sus seguidores), a su vez, por otros usuarios de la red. Se ha escogido Twitter pues, al tener un formato de fragmentos cortos de texto, favorece la identificación de temas concretos y resulta menos complicado analizar cómo estas piezas de información (los tweets) se propagan por el grafo de usuarios.

En cuanto a las aplicaciones de estas investigaciones, vemos que estimar la influencia de un nodo en un proceso de difusión tiene importancia, puesto que nos da una forma de cuantificar los patrones de influencia y los roles de usuarios. Los modelos nos permiten predecir los caminos por los que puede expandirse información en redes sociales, identificando los nodos potencialmente influyentes, de manera que podría aplicarse en campos como publicidad y marketing, política, etc. Puede resultar útil cuando se desea expandir cierta información con rapidez y llegando al mayor número de receptores posible.

1.2 Objetivos

El objetivo principal que se persigue en este trabajo es estudiar la difusión de información mediante el análisis de estructuras de cascada.

Los objetivos específicos son:

- Obtener datos de Twitter en tiempo real, haciendo uso de la API de Twitter.
- Extraer cascadas de información a partir de los datos obtenidos.
- Inferir los caminos posibles con modelos de influencia.
- Simular su acción y ver qué tal reproducen los caminos observando los datos.
- Sacar conclusiones sobre qué modelo se ajusta más a la realidad.
- Analizar las características de la red mediante métricas para obtener patrones y datos de utilidad.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1 – Introducción.** Se presenta el problema que ha motivado el desarrollo de este trabajo y los objetivos que se quieren alcanzar.
- **Capítulo 2 – Estado del arte.** Se exponen los estudios posteriores que han servido de base de este proyecto, de manera que se pone en contexto el trabajo realizado.
- **Capítulo 3 – Construcción y análisis de cascadas**. Corresponde a la primera fase del desarrollo. Primero se profundiza en las cascadas de información, cómo son y cómo se obtienen. Después de detallan los experimentos llevados a cabo y los resultados obtenidos.
- **Capítulo 4 – Modelos de influencia y simulación**. Corresponde a la segunda fase del desarrollo. Primero se profundiza en los modelos de influencia, sus características y tipos. Después se describe el proceso de simulación llevado a cabo y sus resultados.
- **Capítulo 5 – Conclusiones y trabajo futuro**. Se exponen las conclusiones del trabajo y las posibles mejoras futuras y estudios que podrían llevarse a cabo.

2 Estado del arte

Una red social, tal y como describió Newman (2003), es un conjunto de personas con algún patrón de contactos o interacciones entre ellos. La primera vez que fue empleado el término "red social" fue por Barnes (1954), en el contexto de relaciones humanas, pero fue a finales del siglo XIX cuando comenzó a formarse este concepto más en profundidad. Hasta entonces, estas redes eran estudiadas en campos como la sociología, biología, etc, como es el ejemplo de Tönnies (1887). Fue a principios del siglo XXI cuando nacieron las redes sociales online o en línea que hoy conocemos y comenzó una constante evolución de estas que conllevó la creación de diversas plataformas, tales como Facebook (2004), Twitter (2006) o Instagram (2010).

Las redes sociales han supuesto una revolución en la forma de obtener información. En el pasado, las personas tenían como fuentes principales de información el boca a boca, los *mass media* (televisión, radio, periódicos, etc) y de alguna manera tenían la posibilidad de acceder a algún tipo de información más específica (por ejemplo, proporcionada por una revista especializada en un tema concreto), aunque no resultaba del todo sencillo encontrar exactamente la información deseada. Sin embargo, con la llegada de las redes sociales, la fuente de información ha crecido de forma desmesurada y resulta mucho más fácil acceder a información muy concreta de casi cualquier campo. Además, los usuarios de estas redes han obtenido una gran capacidad de comunicar, es decir, traspasar información a otros individuos, lo que antes sólo les permitía el boca a boca y abarcaba un público mucho más limitado. Cuando una pieza de información se distribuye desde un individuo o comunidad hacia otros, se está produciendo un proceso de **difusión de información** (también llamado de propagación o diseminación), tal y como se describe en estudio de Li, M. et al. (2017). En este trabajo se tratará concretamente la difusión de información en redes sociales.

Actualmente, las redes sociales juegan un papel imprescindible en la difusión de información a gran escala, pues proporcionan acceso a cientos de millones de usuarios de Internet en todo el mundo a una extensa fuente de datos, así como les permite producir su propio contenido. Han demostrado ser cruciales en muchas situaciones, como por ejemplo Twitter durante las elecciones presidenciales de 2008 (Hughes y Palen, 2009).

Dado el impacto de las redes online en la sociedad, se han realizado muchos estudios con la finalidad de entender este fenómeno para poder darle uso y extraer información valiosa. Eventos, problemas, conflictos entre ideologías, etc, suceden y evolucionan muy rápidamente en las redes sociales; y su captura, comprensión, visualización y predicción se están convirtiendo en objeto de interés tanto para los usuarios finales como para los investigadores. Como se expone en el trabajo de Guille et al. (2013), esto está motivado por el hecho de que comprender la dinámica de estas redes puede ayudar a mejorar la difusión de eventos futuros (por ejemplo, analizando estallidos de olas de información), resolver problemas (por ejemplo, prevenir ataques terroristas, anticipar peligros naturales), optimizar el rendimiento comercial (por ejemplo, optimizar campañas de marketing social). etc. Por lo tanto, se han desarrollado en los últimos años una variedad de técnicas y modelos que sirven para capturar la difusión de la información en las redes sociales online, analizarla, extraer conocimiento de ella y predecirla e incluso tratar de controlar este fenómeno. En este trabajo se van a emplear algunas de esas técnicas, que más adelante serán explicadas, y se van a incluir algunas nuevas con el fin de expandir el conocimiento que se tiene de la propagación de datos en redes sociales online.

El estudio de la difusión de información en las redes sociales online plantea las siguientes preguntas:

- ¿Qué miembros de la red juegan un papel importante en el proceso de difusión?
- ¿Cómo, por qué y a través de qué caminos la información se propaga, y se propagará en el futuro?
- ¿Qué piezas de información o temas son populares y se difunden más?

En base a estas preguntas, se desarrollan tres temas englobados dentro del estudio de la difusión de información: identificación de focos de influencia, modelado de difusión de información y detección de temas populares.

La identificación de focos de influencia y el modelado de la difusión de información se encuentran fuertemente relacionados entre sí, puesto que dentro del modelado encontramos el grupo de los modelos de influencia, que tienen un papel muy importante en este trabajo. Por ello, en la siguiente sección se tratarán los modelos de difusión y la detección de influencia será uno de los temas principales de los que se hablará.

En cuanto al tercer punto, hay una parte del estudio en la que se eligen los grupos de datos sobre los que se van a realizar los experimentos y, para la elección de dichos datos, se tiene en cuenta el tema de la información que se recoge y se intentará ver las diferencias de los resultados que puedan tener que ver con él (en el Capítulo 3 se detalla este aspecto). Sin embargo, no se utilizará la detección de temas populares como tal, por lo que no será tratada en una sección de este capítulo.

2.1 Modelado de difusión de información

Muchas investigaciones, como por ejemplo las de Christakis y Fowler (2007) o Zhang y Wu (2012), han tratado de analizar la difusión de información, intentando encontrar qué factores afectan al proceso, por qué algunas piezas de información se difunden más rápido que otras y cómo se produce esta propagación. Para ello, se ha utilizado el método de los modelos, que son de gran importancia a la hora de entender este fenómeno.

El método de los modelos o modelado consiste en realizar una representación conceptual de un fenómeno que ocurre en la realidad con el fin de describirlo, reproducirlo o explicarlo. Un ejemplo de modelo pueden ser las fórmulas de la ley de gravitación universal de Newton (1687), que tratan de describir el comportamiento de los cuerpos frente a la gravedad. En este caso el fenómeno que quiere describirse para poder ser explicado y analizado es el de la difusión de información en las redes sociales; más adelante se entrará en detalle en algunos modelos de este proceso.

Como describen Guille et al. (2013), el proceso de difusión está caracterizado por dos aspectos: su estructura, es decir, el grafo que representa los caminos tomados por la información y quién influencia a quién en la decisión de continuar con la difusión; y su dinámica temporal, es decir, la evolución de la tasa de difusión (valor obtenido por la cantidad de nodos que adoptan piezas de información a lo largo del tiempo). Una forma simple de explicar el proceso de propagación es considerar que un nodo puede estar activado (ha recibido una pieza de información y trata de propagarla) o no. Así, la propagación puede verse como la sucesiva activación de nodos a través de la red.

Normalmente, los modelos desarrollados asumen que los usuarios sólo pueden ser influenciados por acciones que captan por sus conexiones. De esta manera las secuencias de activación pueden sacarse según los tiempos en los que fueron publicados los mensajes. Pero esto no proporciona la información completa, por ello, se necesitan modelos que predigan el mecanismo subyacente de difusión. Podemos distinguir dos categorías de modelos: modelos explicativos y modelos predictivos. Esta clasificación está basada en la propuesta por Li, M. et al. (2017).

2.1.1 Modelos explicativos

Los modelos explicativos tienen como objetivo examinar los procesos de difusión de información e identificar los factores que afectan para poder entender mejor este fenómeno. Permiten volver a trazar el camino seguido por la pieza de información y son útiles para comprender cómo la información se ha propagado. Estos caminos que sigue la información se explorarán en profundidad en este trabajo, sobre todo en el Capítulo 3, donde serán denominados cascadas de información.

Este tipo de modelos van a ser clasificados en dos grupos: modelos epidémicos y modelos de influencia.

Modelos epidémicos

Los modelos epidémicos son empleados en el campo de la biología para, a través de ecuaciones diferenciales, intentar dar con una representación de la realidad lo más aproximada posible. En concreto, se representa la propagación de una enfermedad en una población.

Los procesos de difusión de información pueden ser considerados como un proceso de propagación epidémica, en el que hay usuarios infectados y usuarios que son susceptibles a un patógeno. Del mismo modo en que el virus puede ser expandido desde los infectados a los susceptibles, la información puede ser difundida desde los comunicadores a los receptores en una red social. Es por ello que resulta interesante describir de forma superficial los modelos epidémicos, a pesar de que este trabajo no centre su atención en ellos.

El modelo más básico de este grupo es el modelo SI, propuesto por Pastorsatorras (2001); fue llamado así por hacer una distinción de dos grandes grupos en la población: susceptibles (S) e infectados (I). En él se representa cómo los individuos susceptibles pueden pasar a estar infectados.

A partir del modelo SI, han ido desarrollándose nuevos modelos aumentando su complejidad añadiendo nuevos cambios de estado del individuo, como en el modelo SIS (Newman, 2003), donde los individuos infectados podrían curarse y volver a pertenecer al grupo de susceptibles; o introduciendo nuevos grupos en la población, como el modelo SIR (Liu, D. et al., 2014), donde se tienen en cuenta los individuos recuperados (R), que una vez curados no podrán expandir de nuevo el virus.

Muchas investigaciones de difusión de información en redes sociales están basadas en estos modelos clásicos, debido a las similitudes entre ambos procesos de propagación, más adelante se da un ejemplo de ello. Sin embargo, existen algunas diferencias, puesto que la difusión de información está relacionada con el tiempo, la fuerza de conexión entre usuario, el contexto de la información, factores sociales, la estructura de la red, etc. Por ello, se han

desarrollado nuevos modelos, añadiendo modificaciones que se adapten a las condiciones mencionadas.

Un ejemplo de modelo epidémico en red social es el modelo SEIR (Wang, C. et al., 2014), en el que se añadieron los nodos expuestos (E) al modelo clásico SIR. Con él se analizaban los factores de frecuencia de inicio de sesión de usuarios y el número de amigos de los usuarios. Sus resultados demostraron que la frecuencia de inicio de sesión es directamente proporcional a la velocidad y el rango de transmisión de información.

Modelos de influencia

La difusión de información en redes sociales puede entenderse a través del análisis de la influencia. Como ya se ha comentado, la difusión está en parte caracterizada por su estructura, que se basa en un conjunto de nodos que transmiten y reciben información mediante la influencia que ejercen unos sobre otros. Si se es capaz de determinar cómo actúa esta influencia, se podrán obtener muchas propiedades de los procesos de propagación.

Los modelos basados en influencia pueden dividirse en dos categorías: influencia individual e influencia de comunidades. Aunque en este trabajo únicamente se han empleado modelos basados en influencia individual, se comentarán ambas categorías para poner en contexto las investigaciones relativas a la influencia.

- **Influencia individual.**
  Se refiere a la investigación relacionada con los líderes de opinión o focos de influencia. Estos son nodos que pueden jugar un rol de "puente" en la difusión de información, de manera que la información pasa de unos nodos a otros a través de ellos, por lo que tienen cierta influencia sobre otros usuarios en la red.

  Para cuantificar la influencia que tienen estos nodos generalmente se han usado métricas de centralidad, como la de *closeness* (da valor a los nodos según su posición en el grafo; será más alto cuanto más cercana sea al resto de nodos en promedio) u otros algoritmos como PageRank. En otros casos también se han tenido en cuenta otros factores para que la precisión aumente, como puede ser el comportamiento de usuario o su actividad en la red, aunque puede dar cierta subjetividad al proceso y son más difíciles de determinar.

  Un ejemplo concreto de esto es el modelo que propusieron Chenxu et al. (2015), que se trataba de un método para modelar y medir la influencia basándose únicamente en la estructura de la red, en el que el proceso de difusión se describía con un grafo dinámico dirigido.

  Jiaxin et al. (2014) propusieron un método para medir la influencia social prediciendo la habilidad de un usuario de diseminar información. En concreto, la evaluación de influencia se basaba en el recuento de retweets.

  En este trabajo se ha trabajado con estas ideas para implementar modelos de influencia, que serán explicados con detalle en el Capítulo 4.

- **Influencia de la comunidad.**
  En general, una comunidad es un grupo de personas con algunas propiedades comunes; en redes sociales, puede verse como un subconjunto de la red donde los

usuarios están densamente conectados y tienen atributos similares, por ejemplo, algún interés común. Aunque la estructura de la red social cambie a lo largo del tiempo, las comunidades se mantienen relativamente estables, y pueden resultar un gran foco de influencia en la red en la que se encuentran. El principal reto en este campo es detectar esas comunidades y muchos métodos han sido propuestos para este fin, como es el caso de Yang et al. (2014).

La detección de comunidades es un área de gran importancia en el ámbito de las redes sociales, pues tiene aplicaciones tales como detectar posibles comunidades influyentes con ideologías peligrosas, o para hacer sondeos en política. Sin embargo, este trabajo no abarca este campo de modelos de influencia por lo que no se profundizará más en ello.

2.1.2 Modelos predictivos

Los modelos predictivos pretenden predecir cómo un proceso de difusión específico se desplegaría en una red dada, mediante el estudio de trazas de difusión. A diferencia de los modelos explicativos, no se limitan a estudiar los datos y sus características presentes, sino que van más allá, estudiando lo que podría llegar a ocurrir en el futuro.

En una red social, cuando una pieza de información es publicada por un usuario, la información será propagada rápidamente por la red y sería útil poder predecir cómo ocurrirá este proceso. Por ejemplo, en el caso de la expansión de información negativa para una población, como es el caso de bulos o ideologías peligrosas o violentas, sería de gran ayuda ser capaces de prever estos procesos para poder combatirlos. Los modelos predictivos se usan para predecir los futuros procesos de difusión de información es redes sociales basándose en ciertos factores.

En esta categoría hay dos modelos importantes llamados modelo de Cascadas Independientes (IC, Independent Cascades; Goldenberg et al., 2001) y modelo de Umbral Lineal (LT, Linear Threshold; Granovetter, 1978). Ambos modelos asumen que existe una estructura de grafo estático (es decir, que asume que no se añaden nuevos enlaces ni nodos a la red) subyacente en la difusión y se centran en ella.

Anteriormente se ha mencionado que la propagación de información puede verse como un proceso de activación de nodos, en el que los nodos se activan cuando la información llega hasta ellos. Estos modelos traducen este proceso en un grafo dirigido donde cada nodo puede ser activado o no, pero una vez activado no puede volver a ser desactivado. Para ambos modelos, el proceso de difusión es iterativo y síncrono en un eje de tiempo discreto, partiendo de un grupo inicial de nodos activos.

El modelo IC requiere asociar una probabilidad de difusión a cada enlace del grafo y, para cada iteración, los nodos recién activados intentan activar a sus vecinos una vez, con la probabilidad definida en el enlace que les une. Por su parte, el modelo LT define un grado de influencia en cada enlace y un umbral de influencia en cada nodo y, en cada iteración, los nodos inactivos son activados por sus vecinos activos si la suma de los grados de influencia excede su propio umbral de influencia.

En ambos modelos, el proceso acaba cuando no hay nuevas transmisiones posibles, es decir, no se puede activar ningún nodo vecino. Una diferencia existente entre los dos mecanismos es que reflejan dos puntos de vista diferentes: IC se centra en el emisor y LT en el receptor,

puesto que en IC es el emisor quien intenta activar a sus vecinos mientras que en LT son los receptores los que tratan de captar la información. A lo largo del tiempo, se han utilizado estos modelos para las investigaciones de difusión de información y se han realizado modificaciones sobre ellos para obtener mejoras, como por ejemplo, hacer que el proceso pase a ser asíncrono.

En este trabajo se han utilizado las ideas propuestas por varios de los modelos explicados en este apartado 2.1 con algunos cambios. Por un lado, se han recuperado las cascadas propuestas por los modelos explicativos, que definirán la estructura de los procesos de propagación (se explica con más detalle en el Capítulo 3). Por otro lado, se ha utilizado la idea del proceso iterativo síncrono y sobre un eje de tiempo discreto de los modelos predictivos para crear un programa de simulación de transmisión de piezas de información (se explica el algoritmo más en detalle en el Capítulo 4). En esta simulación se usa la idea del modelo IC de asignar una probabilidad y utilizarla para decidir si la información se transmite o no, salvo que esta se le asigna a cada nodo en lugar de al enlace.

3 Construcción y análisis de cascadas

Este apartado corresponde a la primera fase del desarrollo del proyecto. Esta se centra en la construcción de cascadas de información a partir de datos recogidos en Twitter y su posterior análisis.

En primer lugar, se explicará el concepto de cascada y se describirán los criterios seguidos para su construcción, así como el algoritmo utilizado. Después, se describirán los experimentos llevados a cabo para el análisis de las cascadas y los resultados obtenidos.

3.1 Cascadas de información

El proceso de difusión de información está caracterizado por su estructura, es decir el grafo que transcribe quién influencia a quién en la propagación. A esta estructura la llamaremos cascada.

Definición 3.1 Cascada. Una cascada se define como el conjunto de nodos y enlaces que representan la difusión de una pieza de información concreta. Hay un nodo raíz, que corresponde con el usuario que publicó la información original. Cada uno de los demás nodos corresponde con un usuario al que le ha sido transferida la información y la han propagado. Las cascadas tienen estructura de grafos dirigidos acíclicos, donde los enlaces se dirigen desde un nodo hasta otro y representan que en esa dirección fue propagada la información.

Para que haya cascadas de información, es necesario que exista un **grafo de "follows" o de seguidores**, lo que significa que los nodos representarían a los usuarios de una red y los enlaces dirigidos equivalen a la relación "usuario sigue a usuario". Esta relación de "seguimiento" unidireccional se produce en muchas redes sociales, como es el caso de Twitter, donde seguir a alguien equivale a poder acceder a la información que éste publica. Esta relación genera dos estados en los nodos: un nodo n1 puede ser **seguidor** (o follower) de un nodo n2, en cuyo caso habría un enlace desde n1 hasta n2 en el grafo de "follows"; y un nodo n1 puede ser **amigo** (o friend) de n2, donde el enlace iría desde n2 hasta n1. Por tanto, si n1 es seguidor de n2, quiere decir además que n2 es amigo de n1.

Las cascadas de información se producen por las interacciones (propagar información) que se dan entre los nodos de estos grafos de "follows", los cuales son en realidad un subconjunto o subgrafo del grafo total que generan los usuarios de una red social.

En la Figura 3-1 se muestran dos grafos dirigidos. El primero corresponde a un grafo de follows de una red y el segundo a una cascada de información. En ella, el nodo raíz sería u1, el que publicó originalmente la pieza de información; el resto de nodos la han recibido y publicado para que otros nodos puedan recibirla también. En este ejemplo, u2 y u3 reciben la información de u1, según los enlaces dirigidos. En el caso de u4, dado que es seguidor tanto de u2 como de u3, la información ha podido ser recibida desde ambos, por lo que en la cascada se representan los dos enlaces.

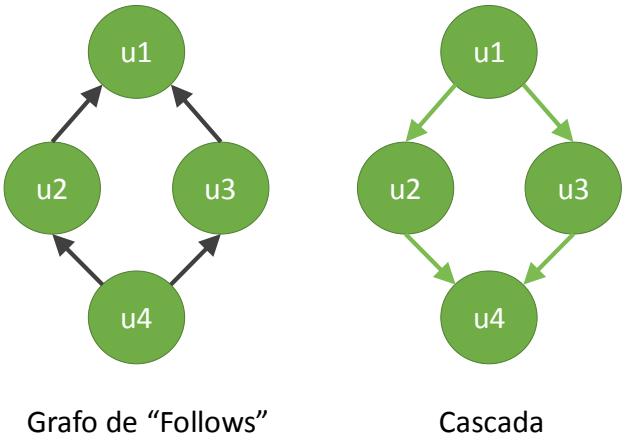

Figura 3-1 Ejemplo de grafo de “follows” y cascada

3.1.1 Reconstrucción de cascadas

En el trabajo se ha utilizado la red social Twitter para recoger cascadas de información producidas por las interacciones de usuarios en esta red. En ella, los usuarios pueden tener relaciones de seguimiento y de amistad, como se ha explicado en el apartado anterior; se pueden publicar piezas de información llamadas tweets y los demás usuarios pueden acceder a ellos y retuitearlos, es decir, publicar ese mismo tweet (de esta manera se propaga la información). Cada usuario de Twitter tiene un timeline donde aparecen los tweets publicados por sus “amigos”, por lo que tienen acceso a esa información de manera sencilla.

El API pública de Twitter no facilita información detallada sobre qué camino exacto siguen las cadenas de retweet. Por ello, se ha desarrollado un proceso de reconstrucción de cascadas, en el que se identifican los posibles caminos que han seguido los tweets.

Con este fin, se han procesado unos conjuntos de datos recuperados de Twitter. Estos datos recuperados aportan la siguiente información: cuál es el tweet original, el usuario que lo publicó y el momento en el que se creó; así como los usuarios que retuitearon dicho tweet y el momento en el que lo hicieron. A partir de cada tweet original, se ha obtenido una cascada: un grafo que representa los posibles caminos que ha seguido la información entre usuarios.

Para reconstruir las cascadas, se parte del grafo de follows recuperado de Twitter, que es un subgrafo de la red social completa. Los usuarios de este grafo propagan información entre ellos y estas interacciones son las que generan las cascadas, tal y como se ha explicado anteriormente. Para establecer un enlace en la cascada, el protocolo que se ha establecido consiste en que la fecha de publicación del tweet debe ser anterior en el nodo padre que en el hijo (el nodo padre tubo que publicar el tweet antes para que el hijo lo pueda retuitear posteriormente) y debe existir una relación de seguimiento (el hijo sigue al padre).

Sin embargos, en la obtención de datos de Twitter se daban casos en los que un usuario retuiteaba un tweet sin mantener una relación de seguimiento con ningún otro usuario que hubiese publicado el tweet antes que él. En un primer momento se planteó descartar este tipo de nodos, dado que se consideró una situación poco probable que un usuario retuitease a alguien que no siguiera, es decir, algo fuera de su timeline. Sin embargo, la posibilidad de que los usuarios llegasen a una cuenta mediante búsqueda, o incluso en los tweets destacados (*trending topic*, opción que da la plataforma para ver los tweets más destacados del

momento), no parecía algo tan improbable, por lo que se decidió recoger en la cascada estos casos.

La manera de recoger en la cascada estos usuarios que retuitean sin mantener relación de amistad con otro usuario que haya publicado esa información anteriormente, es añadirlos como nodos aislados, es decir, un vértice del grafo sin ningún enlace asociado.

Por todo lo explicado, las cascadas presentan una estructura de grafo dirigido acíclico. Debe ser dirigido pues la dirección del enlace representa hacia donde se ha propagado la información, y no pueden producirse ciclos (caminos que empiezan y acaban en el mismo nodo) porque si hubiese un enlace entrante en un nodo que presenta un enlace saliente supondría una contradicción con la fecha de publicación de los tweets.

A continuación, se detalla el algoritmo desarrollado de construcción de cascadas de información. Se ha de aclarar que este algoritmo construye cascadas cuyos nodos son tweets en lugar de usuarios; sin embargo, esto es equivalente a las cascadas de usuarios dado que existe una correspondencia única entre tweet y usuario. Por ello, una vez construida la cascada de tweets resulta sencillo obtener la cascada de usuarios sustituyendo los nodos según la correspondencia entre usuario-tweet.

El algoritmo sigue los pasos siguientes:

1. Insertar el tweet raíz como vértice de la cascada y el usuario que lo publicó en "Usuarios explorados".
2. Insertar los retweets del tweet raíz en una cola de prioridad, que los ordena por fecha de publicación de forma descendente.
3. Extraer el primer tweet de la cola e insertarlo como vértice de la cascada.
4. Intersecar los "Usuarios explorados" con los "Amigos" del usuario que se está explorando.
5. Crear un enlace desde los tweets publicados por los usuarios pertenecientes a la intersección, hasta el tweet que se está explorando.
6. Repetir desde el punto 3 hasta que la cola esté vacía.

La Figura 3-2 ilustra el algoritmo con un ejemplo sencillo. En el paso (1), se inserta en la cascada como nodo raíz el tweet t0, el tweet original. Sus retweets se insertan en la cola de prioridad, ordenados por fecha de publicación de forma descendente; de manera que los tweets más antiguos, es decir, más próximos en el tiempo al tweet original, serán los primeros en salir de la cola.

En los pasos (2), (3) y (4), el procedimiento consiste en extraer un tweet de la cola y comprobar la intersección entre los usuarios de los tweets que se han explorado hasta el momento y los "amigos" del usuario que se está explorando. Entonces, se inserta el tweet en la cascada y se le añaden enlaces entrantes desde los tweets de los usuarios obtenidos, indicando así que se trata de un posible retweet directo de estos.

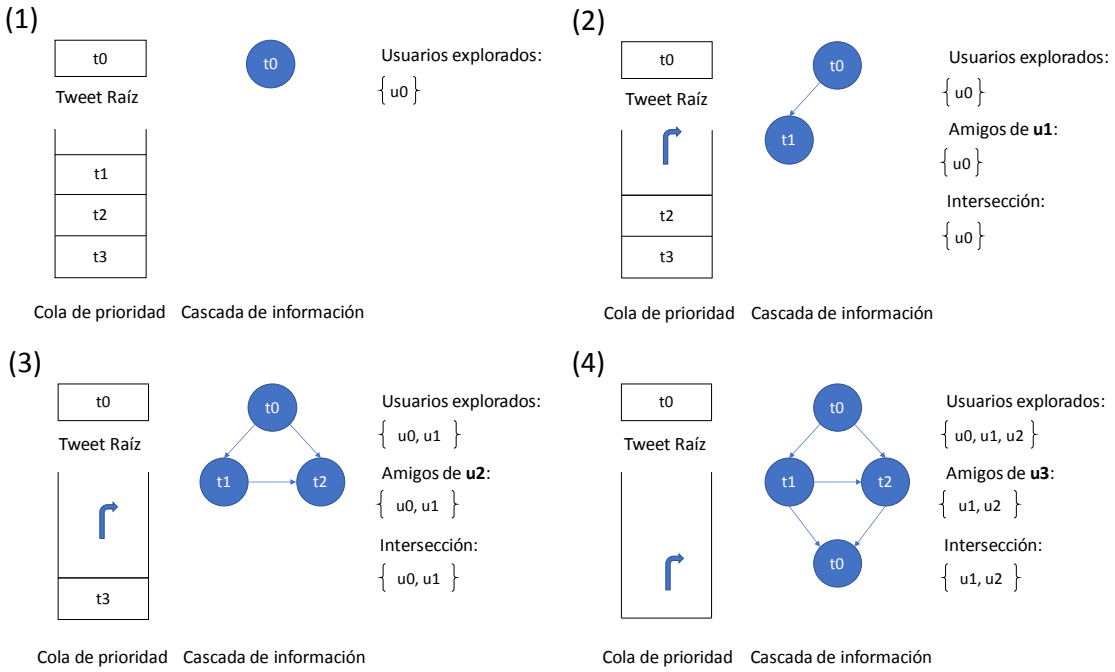

Figura 3-2 Ilustración de la reconstrucción de cascadas.

3.1.2 Análisis topológico

Se ha realizado un análisis de las cascadas reproducidas a partir de los datos de Twitter, con el fin de estudiar las características de la difusión de información y poder obtener información de este fenómeno.

Para este análisis se han utilizado métricas de caracterización de grafos, que han sido desarrolladas con ayuda de la librería JUNG. Todas ellas son métricas básicas del campo del análisis de redes sociales, salvo las métricas CR y RFR, que se han obtenido del trabajo de Taxidou y Fischer (2014). A continuación, se dará una breve explicación de cada una de ellas.

Definiremos para las fórmulas de las métricas una cascada de información $C = (U, E)$, con U siendo el conjunto de nodos y E el conjunto de enlaces, donde $u_r \epsilon U$ es el nodo raíz.

- **Tamaño**. Se medirá el tamaño de la cascada por su número de nodos.
$$Tamaño = |U|$$
El estudio de esta métrica es interesante puesto que indica cuánta cantidad de usuarios ha participado en el proceso de difusión y, por consiguiente, qué alcance ha tenido cada pieza de información.
- **Número de componentes conexas**. Se calcula el número de componentes débilmente conexas, esto es, si eliminamos la dirección de los enlaces, todos los pares de nodos de la componente son mutuamente accesibles.
En la Figura 3-3 se muestra un grafo dirigido, que bien podría tratarse de una cascada de información, y pueden distinguirse en tres colores las tres componentes conexas por la que está formado. En cada una de ellas, si se eliminasen la dirección de los enlaces, podría trazarse un camino entre cualquier par de nodos de la componente.

- **Número de componentes conexas sin contar nodos aislado**. Se realiza el mismo cálculo que en el punto anterior, pero no se cuentan aquellas componentes formadas por un único nodo aislado.
  En la Figura 3-3, de las tres componentes conexas que se observan diferenciadas por colores, esta métrica contaría únicamente dos, puesto que descartaría la componente amarilla formada por un nodo aislado, un vértice sin enlaces.

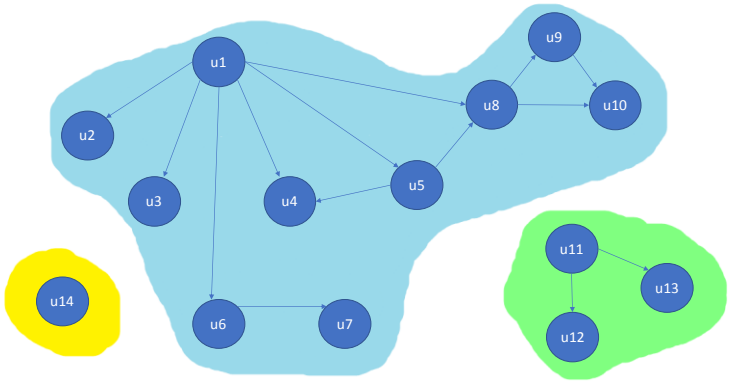

Figura 3-3 Grafo con tres componentes conexas distinguidas por colores.

- **CR (Conectivity Rate)**. Determina el porcentaje de usuario que tienen al menos una conexión con otro usuario.

$$CR = \frac{|\{u \in U | (u', u) \in E \ \vee (u, u') \in E\}|}{|U|}$$

- **RFR (Root-Fragment-Rate)**. Porcentaje de usuarios conectados con el nodo raíz, es decir, con el usuario que publicó el tweet original; ya sea directa o indirectamente.

$$RFR = \frac{\left|\{u_j \in U \mid sii\ existe\ un\ camino\ u_r, .., u_j\ en\ C\}\right|}{|U|}$$

  Para explicar mejor las métricas CR (Conectivity Rate) y RFR (Root-Fragment-Rate), con un sencillo ejemplo veremos su significado. En la Figura 3-4 se muestra una cascada con tres componentes conexas: la componente del nodo raíz u1; la de los nodos u11, u12 y u13; y la del nodo aislado u14. Como se ha explicado en apartados anteriores, esta desconexión se debe a que algunos usuarios retuitean la información sin seguir al usuario que publicó el tweet. De esta forma, pueden aparecer nodos aislados o incluso subcascadas con un nuevo nodo raíz, como es el caso de este ejemplo.
  La métrica CR mide el porcentaje de nodos no aislados, es decir, que están conectados con al menos otro nodo (se ignora la dirección del enlace). En el ejemplo la proporción sería de 13 nodos conectados con otro nodo, frente a 14 nodos en total; se obtiene un $CR = 0{,}9285$. La métrica RFR mide el porcentaje de usuarios conectados con la raíz principal directa o indirectamente, es decir, que existe un camino entre dicho usuario y el usuario que publicó el tweet original. En otras palabras, el porcentaje de usuarios que se encuentran en la componente conexa

principal. En el ejemplo de la Figura 3-4 tenemos 10 nodos en dicha componente frente a un total de 14 nodos, de manera que $RFR = 0{,}71429$.

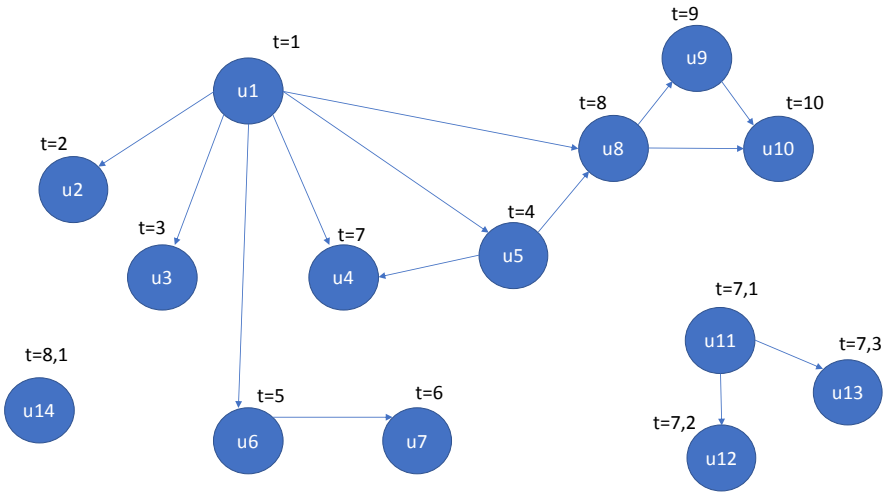

Figura 3-4 Ejemplo de Cascada

Los datos aportados por las métricas CR y RFR junto con el número de componentes conexas son relevantes para el estudio puesto que, si un usuario no presenta ningún enlace entrante en la cascada (exceptuando el raíz), quiere decir que no se puede determinar de quién ha retuiteado la información; no se puede identificar el usuario que ejerció influencia sobre él.

- **Diámetro**. El diámetro es la distancia mínima máxima, es decir, de todas las distancias mínimas entre todos los pares del grafo, se obtiene la mayor. Sólo se tienen en cuenta las distancias entre nodos conectados, es decir, que el valor será el de la componente conexa del grafo con mayor diámetro.
- **Grado promedio**. El grado de un nodo es el número de enlaces en los que participa dicho nodo; en grafos dirigidos, se diferencian el *outdegree*, se cuentan los enlaces salientes, y el *indegree*, se cuentan los entrantes. En la Figura 3-5 se muestra un ejemplo: el *indegree* del nodo u es 2 y del nodo v es 1; el *outdegree* del nodo u es 3 y del nodo v es cero.

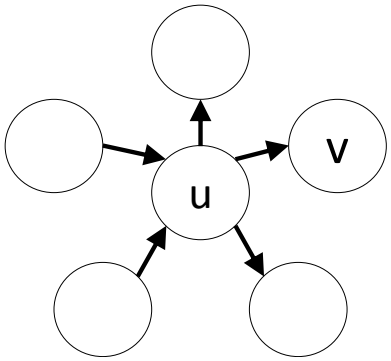

Figura 3-5 Ejemplo de grafo dirigido

En este caso, dado que en las cascadas los enlaces salen del nodo raíz hacia el resto de los nodos (se representa así la dirección en la que se ha difundido el tweet), se calcula el *outdegree* de cada nodo y después se realiza el promedio.

- **Grado promedio sin ceros**. Se realiza el mismo cálculo que el punto anterior, salvo que, cuando el grado de un nodo es cero, este se descarta y no se tiene en cuenta a la hora de promediar.
  Las métricas de grado ayudan a definir la topología de las cascadas. Por este motivo, en este estudio tendrán un papel fundamental a la hora de comparar unas cascadas con otras en la segunda fase del proyecto, descrita en el Capítulo 4.
- **Edad media de retweets**. Es el tiempo promedio desde que el tweet original se publica y es retuiteado. Esta métrica aporta información sobre la velocidad a la que se ha transmitido una pieza de información por la cascada; si la edad media de retweet (tomada en segundos en este trabajo) es baja, quiere decir que la difusión se ha realizado a gran velocidad. Si se combina este dato con el tamaño de la cascada, dará una idea de la viralidad que ha tenido la información (cuánto alcance tiene y en cuánto tiempo). En el apartado de resultados 3.2.2, se desarrolla esta idea.
  La fórmula de la edad media de retweet es la siguiente, siendo t_0 el tiempo de publicación del tweet original y t_i el tiempo de publicación de cada retweet.

$$EMR = \frac{\sum(t_i - t_0)}{Tamaño}$$

3.2 Experimentos

3.2.1 Conjuntos de datos

Como base de todo el desarrollo del proyecto, es necesario un conjunto de datos sobre los cuales realizar los experimentos; en este caso se ha trabajado sobre dos conjuntos formados con datos recuperados de Twitter. Para realizar la recopilación, se ha implementado un programa que accede a los datos de Twitter y los guarda en una base de datos descrita en el Anexo A.

En primer lugar, se han empleado las clases y métodos de la API de Twitter para capturar los tweets en tiempo real filtrándolos por palabras, concretamente se ha usado la API de Streaming[1], la cual cuenta con ciertas limitaciones[2]. Sólo se captura un porcentaje de los tweets y retweets que se van publicando en tiempo real, por lo que es posible que se pierdan alguna "pieza" de la cascada de difusión del tweet. Este aspecto será examinado en los resultados del análisis de cascadas en el apartado 3.2.2, a través de las métricas de conectividad.

Una vez que se han capturado los tweets en la base de datos, junto con la información asociada a ellos (fecha de publicación, usuario que lo publicó, número de seguidores, número de retweets son los más importantes), se procede a recuperar las relaciones de seguimiento entre los usuarios recogidos en base de datos para poder construir un grafo de "follows" sobre el que realizar el proceso de reconstrucción de cascadas. Para ello se han utilizado las clases y métodos ofrecidos por la API de REST[3] de Twitter, que también cuenta con

[1] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[2] https://developer.twitter.com/en/docs/basics/rate-limiting.html

[3] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/overview

limitaciones, por ejemplo, permite realizar un número limitado de llamadas y después bloquea el acceso durante un periodo de tiempo. Sin embargo, estas limitaciones sólo afectan al tiempo de recogida, pero pueden recuperarse todos los datos sin perder ninguno.

El proceso de construcción del grafo de "follows" es el siguiente: se recorren todos los tweets y retweets almacenados en base de datos, se accede al id del usuario que los publicó, a través de llamadas a la API se accede a los "amigos" del usuario y se comprueba si dicho amigo pertenece al conjunto almacenado en la base de datos, en cuyo caso se recoge la relación de amistad. Al finalizar este proceso se cuenta con un grafo que representa las relaciones de seguimiento entre todos los usuarios que han participado en la propagación de los tweets capturados en el proceso anterior.

A continuación, se comentarán los conjuntos de datos recogidos, así como el criterio empleado para la elección de dichos grupos y sus características.

- **Datos con temática de "Cataluña"**. Se ha realizado la recopilación de tweets en tiempo real filtrando la búsqueda por palabras relacionadas con temas de actualidad de Cataluña ("Cataluña", "Puigdemont").
  Filtrando los tweets por temática, lo que se quiere estudiar son las diferencias que pueden darse entre los grafos generados. Esta temática se ha escogido por ser un tema de actualidad, del que mucha gente muy diversa opina; por lo tanto, es muy probable que el grafo de "follows" que se obtenga sea disperso, dado que, al ser un tema tan general, los usuarios no tienen por qué estar relacionados entre sí. En este caso es probable que se obtengan estructuras de tipo estrella, es decir, un gran número de usuarios que no están conectados entre sí, pero que siguen a un mismo usuario popular.
- **Datos con temática de "Operación triunfo"**. Se han recopilado tweets en tiempo real filtrando por palabras relacionadas con club de fans de algunos concursantes ("Almaia", "Aiteda" …).
  En este caso, la temática se ha escogido para dar con un grafo de "follows" más compacto, es decir, que existan muchas conexiones entre muchos usuarios, evitando la estructura de estrella explicada en el conjunto anterior. Por ello, se ha querido buscar un tema más cerrado, del que sólo opinen un grupo de personas concreto. Por esta razón se pensó en un club de fans, pues suele ocurrir que, dentro de este grupo, sus miembros se relacionan entre sí con frecuencia. Sin embargo, se debía elegir un tema de actualidad para poder obtener un gran número de tweets en poco tiempo, por ello se escogió "Operación Triunfo".

Uno de los objetivos es comparar las estructuras generadas por estos dos conjuntos de datos, para estudiar si la diferencia de temas puede afectar a la topología y características y sacar conclusiones sobre ello. En la Tabla 3-1, se muestra un resumen del tamaño de los conjuntos de datos en la que se pueden hacer algunas observaciones de interés. Ambos conjuntos tienen aproximadamente el mismo número de tweets, sin embargo, el número de usuarios es prácticamente la mitad en el segundo conjunto, lo que cabía esperar según las suposiciones que se hicieron: siendo la mitad de usuario se llegan a generar aproximadamente el mismo número de tweets, lo que significa que se realizan muchas más interacciones entre amigos, por lo que parece que el conjunto "OT" es más compacto que el de "Cataluña". Además, el número de cascadas en "OT" es mayor también, lo que da más valor a esta teoría.

El hecho de que las relaciones de follows o de seguimiento sea mucho mayor en "Cataluña" no contradice la teoría, puesto que lo más probable es que se traten de relaciones de seguimiento de multitud de usuarios hacia un único nodo popular en la red, es decir, que se obtendría una estructura de estrella como se ha comentado anteriormente, en lugar de la estructura compacta de muchas conexiones entre todos los nodos.

	Usuarios	Tweets	Retweets	Follows	Nº Cascadas
Cataluña	80.446	285.814	195.031	15.507.948	4630
OT	48.883	292.402	220.906	4.050.005	8115

Tabla 3-1 Descripción de los conjuntos de datos

3.2.2 Resultados

Tras recrear las cascadas, se han procesado mediante métricas y en esta sección se analizan dichos resultados. En las Figuras Figura 3-6 y Figura 3-7, se muestran 9 gráficas por cada conjunto de datos; una gráfica por cada métrica explicada en el apartado 3.1.2.

Si analizamos la primera gráfica, que representa el tamaño de las cascadas, se observa que presenta una distribución "power law". Esto quiere decir que las cascadas de mayor tamaño se dan con poca frecuencia y esta aumenta cuanto más disminuye el tamaño.

En cuanto a las métricas CR (Conectivity Rate) y RFR (Root-Fragment-Rate), teniendo en cuenta la explicación de ambas en el apartado 3.1.2, cuanto más cercano a 1 sea el valor de CR y RFR, más conectividad habrá en la cascada. En las gráficas de CR y RFR de las Figuras Figura 3-6 y Figura 3-7, se observa que la mayor parte de las cascadas tienden a acercarse al valor 1 y, por tanto, la conectividad es bastante alta en general. También se puede corroborar este hecho observando las gráficas correspondientes al número de componentes conexas y número de componentes sin contar nodos aislados (primera gráfica y segunda gráfica de la segunda fila), donde la mayoría de cascadas tienen un número pequeño de componentes, lo que indica bastante conectividad.

Una baja conectividad implica, como se ha comentado anteriormente, que existen nodos en la cascada que no mantienen una relación de amistad con otros nodos de la cascada que hayan publicado la información antes que él. De esta manera, no puede determinarse cuál es el usuario que ha ejercido influencia sobre el nodo en cuestión. Este caso puede darse porque el usuario, en lugar de ver el tweet en su timeline, ha "encontrado" el tweet por otros medios (en la sección de tweets destacados, por ejemplo). Otra opción puede ser que se hayan perdido nodos de la cascada en el proceso de recolección, puesto que, como se ha mencionado en el apartado 3.2.1, la API de Twitter tiene límites en cuanto a la recuperación de datos; este caso sería problemático dado que la cascada se presentaría incompleta. Sin embargo, tras comprobar que los resultados presentan una alta conectividad, se puede concluir que, en general, las cascadas son bastante completas (algo a lo que se ha llegado en otras investigaciones, como la de Taxidou y Fischer, 2014).

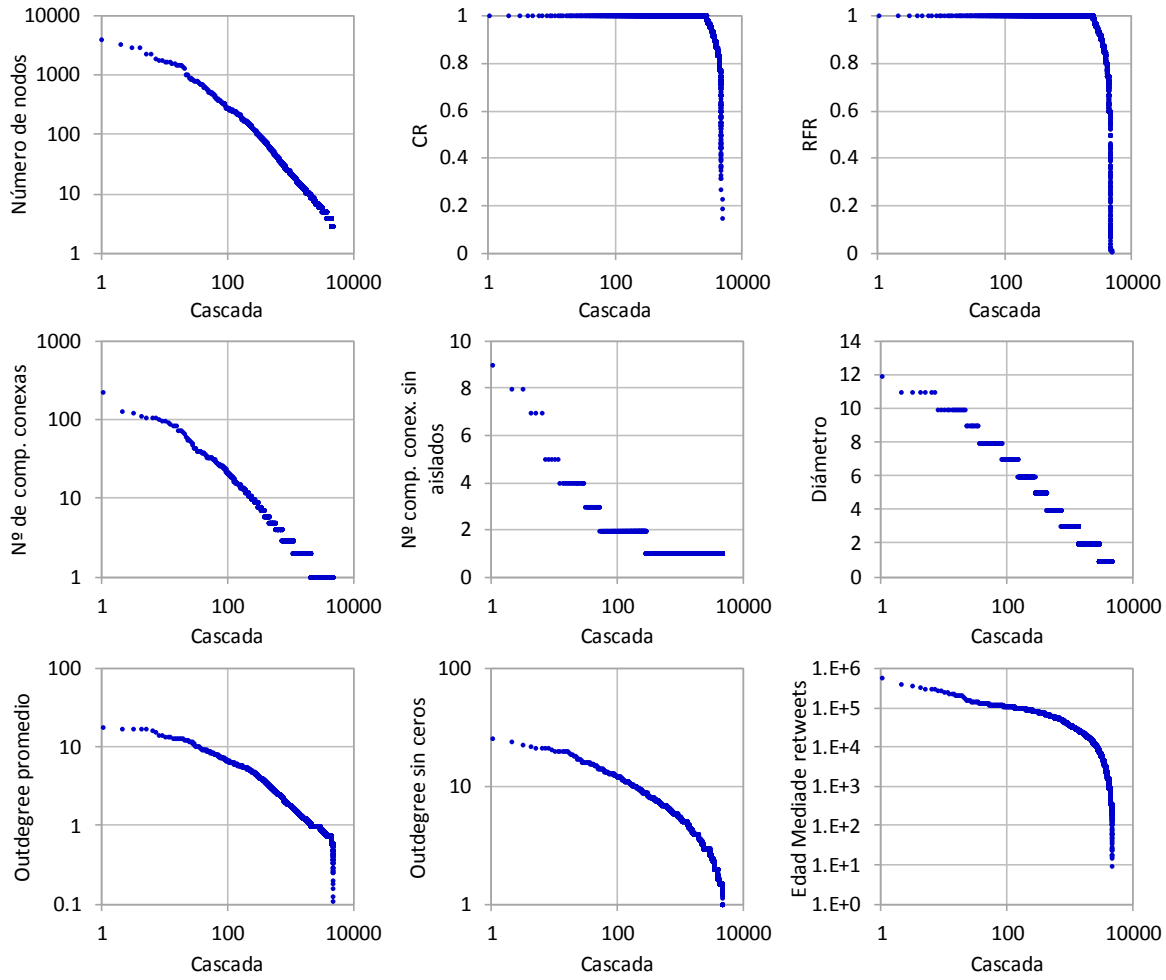

Figura 3-6 Métricas del Conjunto "Cataluña"

Para analizar mejor el diámetro, se ha realizado una gráfica que representa el tamaño frente al diámetro en la Figura 3-8. Se puede observar que presenta una distribución con forma de función logarítmica, es decir, al principio crece muy rápido y después se estabiliza. En cascadas de pequeño hasta mediano tamaño (digamos de 3 a 100 nodos), la relación entre el tamaño y el diámetro se mantiene creciente, lo cual tiene sentido puesto que la información se ha propagado a más usuarios y cabe pensar que se han podido necesitar más "saltos" entre el usuario raíz y los últimos usuarios en retuitear (ha habido nodos entre medias de ambos en el camino de la pieza de información). Sin embargo, en cascadas más grandes (superiores a 100), el diámetro se estabiliza, deja de crecer en función al tamaño.

En cuanto al *outdegree*, se ha calculado de dos maneras: el *outdegree* promedio es la forma convencional, donde se tiene en cuenta el grado de todos los nodos y se hace la media; el *outdegree* promedio sin ceros, no tiene en cuenta los nodos cuyo grado es cero. Con el segundo método se quiere conseguir hallar el grado promedio únicamente de los nodos "influyentes", es decir, los cuales son transmisores en potencia de información. Ambas gráficas presentan también una distribución "power law", como la del tamaño; por tanto, es más frecuente encontrarse un grado promedio más bajo.

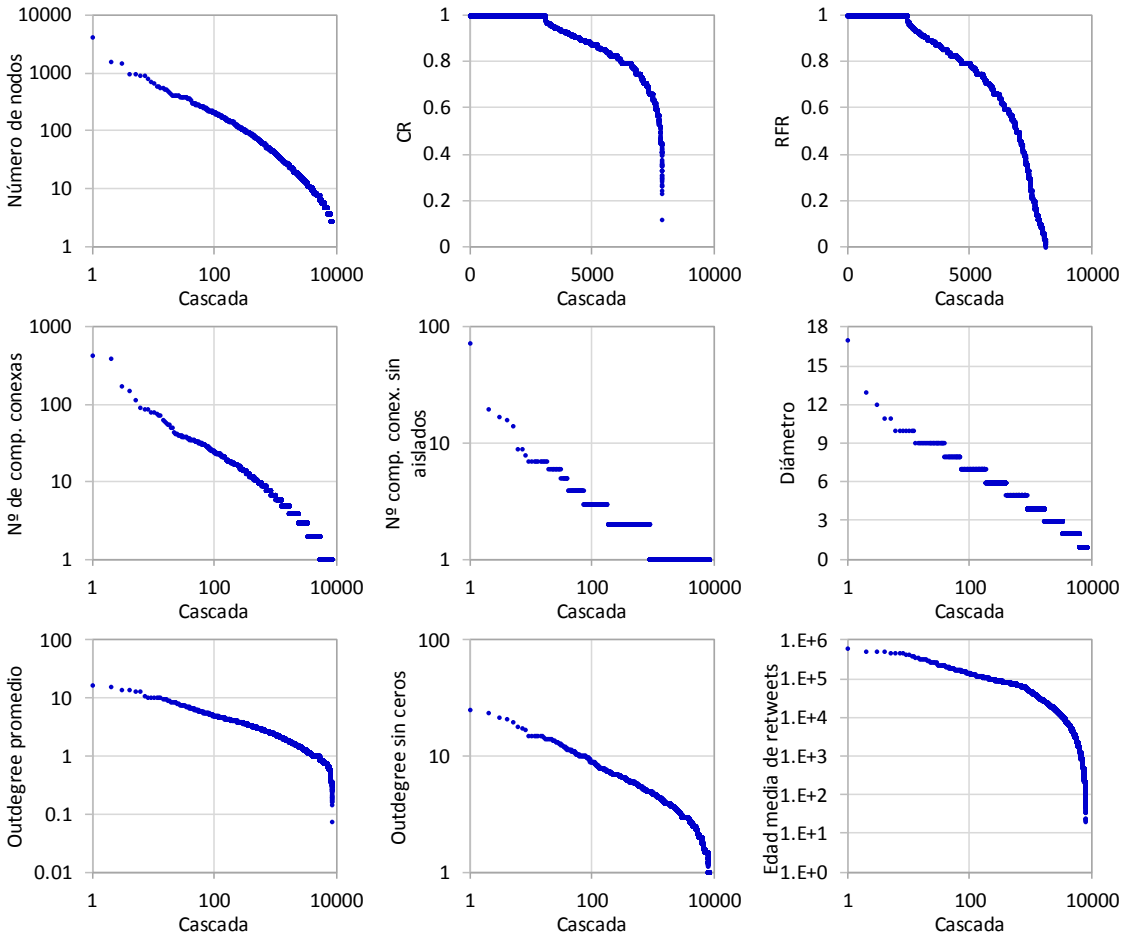

Figura 3-7 Métricas del Conjunto "OT"

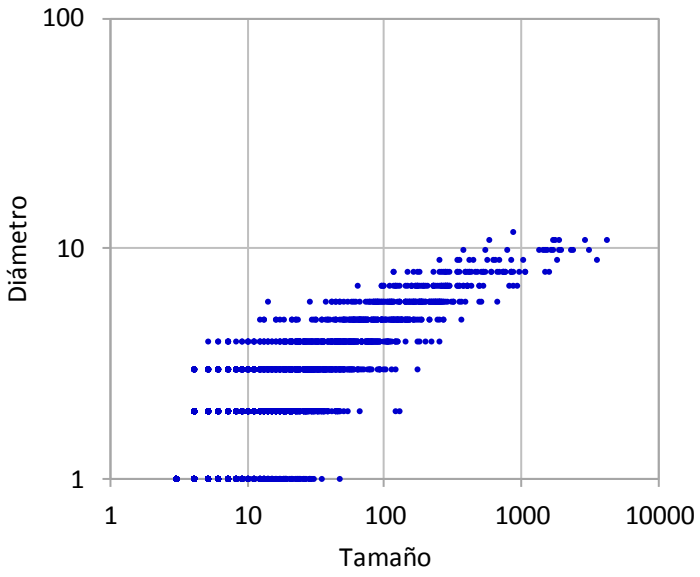

Figura 3-8 Gráfica que relaciona el tamaño con el diámetro en las cascadas del Conjunto "Cataluña"

Se han escogido tres cascadas concretas de tamaños diferentes para hacer una comparación entre ellas. En la Figura 3-9, se observan las tres cascadas representadas como grafos dirigidos, aunque no aparecen los nodos aislados; y en la Tabla 3-2, vemos los valores de las métricas.

Si se observan los valores de la Tabla 3-2, se aprecia que la conectividad es bastante alta en las tres muestras en proporción a su tamaño, lo cual encaja con los datos observados en las gráficas anteriores.

La edad media de retweets, da una idea del tiempo que pasa desde que un tweet se publica hasta que es propagado por otro usuario. En la Tabla 3-2 se ve que cuanto mayor es la cascada mayor es este tiempo. Esto tiene su explicación en que, cuando hay más nodos implicados en la propagación, la difusión del tweet perdura más en tiempo generalmente, hasta que llega a todos los usuarios.

Sin embargo, si se estudia la velocidad a la que se propaga el tweet en cada caso, se puede ver que ésta es mayor en la cascada más grande. Veámoslo paso a paso.

Teniendo en cuenta la fórmula de la edad medio de retweet expuesta en el apartado 3.1.2, la velocidad se calcula de la siguiente manera, siendo $T = máx(t_i - t_0)$. En otras palabras T sería la diferencia del tiempo de publicación del tweet que se publicó en último lugar y el tweet original.

$$Velocidad = \frac{Tamaño}{T}$$

Teniendo esto en cuenta se calculan las velocidades de las tres cascadas de ejemplo, con sus correspondientes tamaños y sus valores T, medidos en horas.

$$Cascada\ pequeña: \quad \frac{5}{246{,}1994} = 0{,}0203\ retweets/h$$

$$Cascada\ mediana: \quad \frac{110}{1078{,}9652} = \ 0{,}1019\ retweets/h$$

$$Cascada\ grande: \quad \frac{1019}{2001{,}0639} = \ 0{,}5092\ retweets/h$$

Como se puede apreciar, el coeficiente más bajo se da en la cascada de mayor tamaño y viceversa, lo que indica que, cuanto mayor es el tamaño la información se propaga proporcionalmente en menor tiempo.

Una posible explicación a este hecho se encuentra en la viralidad de la información. En la tercera cascada la información ha llegado a 1019 usuarios en pocos días, mientras que en la primera cascada el tweet llegó a 5 usuarios en varias horas. Claramente, en la cascada de mayor tamaño se ha producido el fenómeno de viralidad, en el que el tweet ha obtenido una popularidad considerable a corto plazo; en las propagaciones de este tipo suelen intervenir usuarios populares, quizá cuentas verificadas, o que pueden ejercer influencia sobre muchos otros usuarios.

A pesar de que en este ejemplo concreto se produce una relación creciente de la velocidad con respecto al tamaño de la cascada (cuanto más grande es la cascada, a más velocidad se propaga la información), se debe analizar este aspecto de forma general para llegar a una conclusión global. Para ello, se ha representado en la gráfica de la Figura 3-10 la relación entre el tamaño y la velocidad. La línea de tendencia representada en color rojo es creciente, lo que indica que en general sí ocurre que al ser mayor el tamaño de la cascada la velocidad también es mayor.

Sin embargo, en la gráfica se pueden ver representadas cascadas de pequeño tamaño con velocidades altísimas, muy por encima que las cascadas grandes. La razón de estos casos puede darse en que un usuario común publica una información que rápidamente es retuiteada por su grupo de amigos cercanos, siendo una propagación muy rápida; pero estos casos pueden no ser los más interesantes en un estudio de difusión de información, pues en general interesa que el alcance de la propagación sea más extenso.

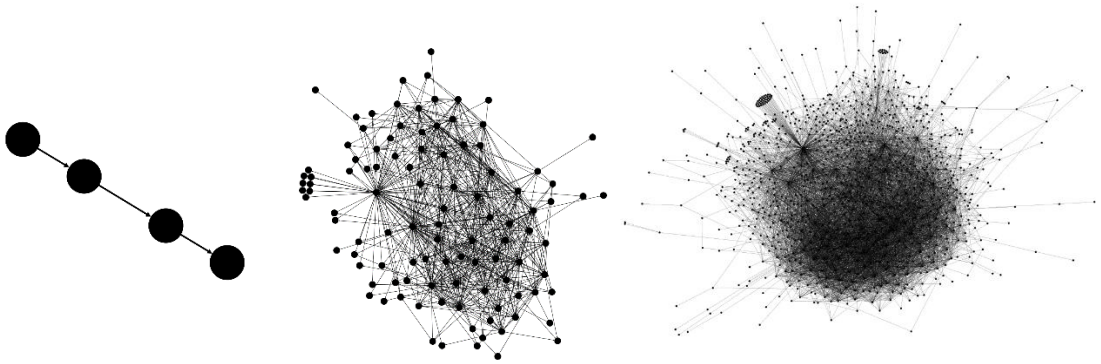

Figura 3-9 Cascadas de distintos tamaños (no se representan los nodos aislados) del Conjunto "Cataluña"

	Cascada Grande	Cascada Mediana	Cascada Pequeña
Tamaño	1019	110	5
CR	0,9195	0,9545	0,8
RFR	0,8763	0,9455	0,8
Nº Comp. Conexas	86	6	2
Nº Comp. Conex. Sin nodos aislados	4	1	1
Diámetro	9	5	3
Outdegree promedio	7,0648	6,1636	0,6
Outdegree promedio sin ceros	11,8600	9,8261	1
Edad media de retweets (s)	133489,417	33471,5596	11709,75
Velocidad (retweets/h)	0,5092	0,1019	0,0203

Tabla 3-2 Métricas de tres cascadas del Conjunto "Cataluña"

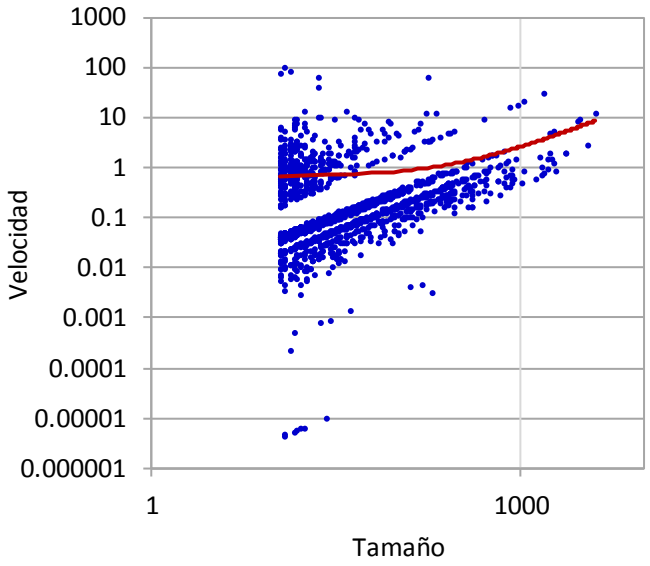

Figura 3-10 Gráfica que relaciona el tamaño con la velocidad en las cascadas del Conjunto "Cataluña", con una línea de tendencia lineal en color rojo

4 Modelos de influencia y simulación

En esta segunda fase del proyecto, se va a ahondar en el concepto de modelo de influencia, el cual nos permitirá refinar los caminos que las cascadas nos daban como posibles, de manera que se pueda caracterizar mejor la red.

Con los modelos de influencia se pretenden entender más en profundidad cómo se ha propagado esa información de forma más precisa. Se abordará el problema de determinar los "caminos de influencia", que expresan la relación de "quién fue influenciado por quién".

Posteriormente, a través de una simulación, se realizará una comparación entre los resultados y las cascadas obtenidas en la primera fase del trabajo, para resaltar los modelos más parecidos a lo observado en la realidad.

4.1 Modelos de Influencia

En el capítulo 3 del trabajo se ha hablado de cascadas de información. En ellas se recogían todos los caminos posibles, bajo una serie de criterios, por los que la información pudo ser difundida. El objetivo de los modelos de influencia es refinar estas cascadas, es decir, de todos los posibles caminos que proporciona la cascada, el modelo de influencia se quedará con uno sólo, el que sea considerado más probable según el criterio del modelo.

Dado que las cascadas eran estructuras de tipo grafo acíclico dirigido, la estructura obtenida tras aplicar un modelo de influencia es un árbol. En la Figura 4-1, podemos ver un ejemplo de lo que podría ser un resultado de aplicar dos modelos diferentes a una misma cascada.

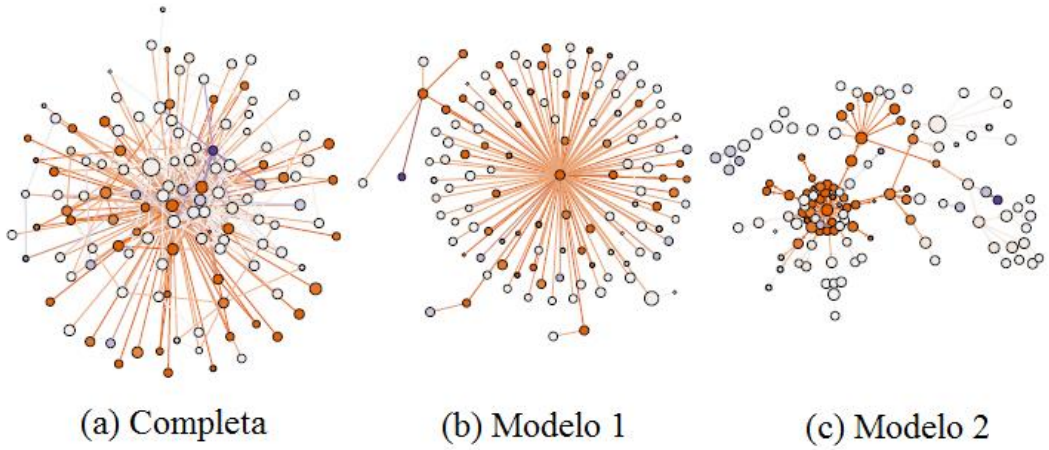

Figura 4-1 Ejemplo de aplicar un modelo de influencia a una cascada. Taxidou y Fischer (2014)

4.1.1 Tipos de modelos

Como se ha comentado, un modelo de influencia necesita un criterio de elección por el cual se escogerá uno de los enlaces entre los usuarios posibles influyentes. Este criterio puede estar relacionado con diversos aspectos de la red. A continuación, se describen los modelos que se han utilizado en el trabajo.

- **Least Recent Influencer (LRI)**: Se basa en que los usuarios están influenciados por la información menos reciente, es decir, la que se publicó antes. En la Figura 4-2 vemos una cascada con algunos de sus enlaces tachados bajo el criterio en el que este modelo se basa, formando un árbol. Por ejemplo, en el nodo u4, nos encontramos con más de un posible camino; según el tiempo de publicación (indicado en la figura con "t = instante de tiempo en el que se publicó el tweet" situada al lado de cada nodo), el usuario que publicó antes el tweet fue u1 y, por tanto, ese es el enlace que prevalece frente al de u5, que es eliminado.

  Este modelo favorece la topología de estrella, es decir, un nodo con numerosos enlaces salientes hacia el resto de nodos. Esto se debe a que el tweet raíz siempre es el que tiene un menor tiempo de publicación por ser el original, por lo que en todos los usuarios que mantengan una relación de seguimiento con el usuario raíz será el enlace que prevalezca de la cascada. En la Figura 4-1, este modelo encaja con el grafo de la imagen (b).

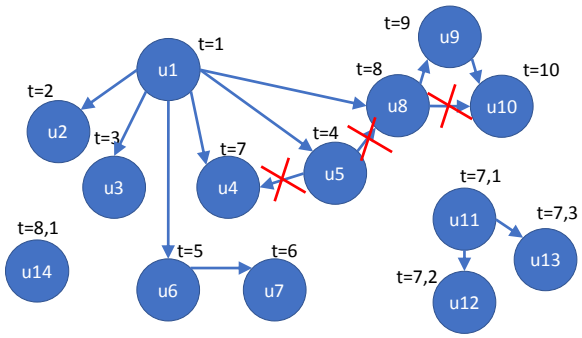

Figura 4-2 Árbol resultante de aplicarle a una cascada el modelo de influencia LRI

- **Most Recent Influencer (MRI)**: Se basa en que los usuarios están influenciados por la información más reciente, la última en publicarse. En la Figura 4-3, se muestra la misma cascada que en el ejemplo anterior con los enlaces tachados que se descartan aplicando este modelo. En este caso, en la elección del camino más probable en el nodo u4, se ha favorecido al usuario u5 por haber publicado la información más tarde que u1.

  La topología que genera este modelo suele ser menos radial que la que se da con el modelo LRI. En la Figura 4-1, el MRI se corresponde con la imagen (c).

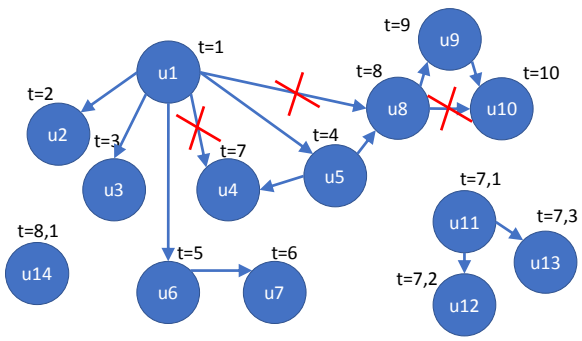

Figura 4-3 Árbol resultante de aplicarle a una cascada el modelo de influencia MRI

- **Most Followered Influencer (MFI)**: Se basa en que los usuarios con más seguidores tienden a ser más populares y, como consecuencia, pueden desencadenar más retweets. En la Figura 4-4, se muestra el número de seguidores de cada usuario (representado como "F = número de seguidores" al lado de cada nodo), y en función a dicho número, el modelo elegirá el enlace que tenga mayor valor. Por ejemplo, el nodo u1 cuenta con 6 seguidores mientras el u5 tiene 2; es por esto que los dos enlaces que llegan a u4, permanece el que lo conecta con u1.

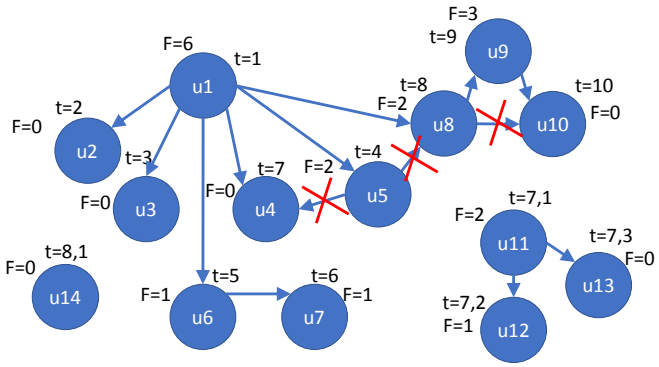

Figura 4-4 Árbol resultante de aplicarle a una cascada el modelo de influencia MFI

- **Most Retweeted Influencer (MRetI)**: Se basa en que los usuarios cuyos tweets han sido más difundidos, emiten contenido interesante para los usuarios de la red y ejercen mayor influencia sobre otros. En la Figura 4-5, el número de retweets conseguidos por los usuarios está representado como "R = número de retweets" al lado de cada nodo. De esta forma, u5 es considerado más influyente que u1 sobre u4, dado que la información que publica es más veces retuiteada en la red.

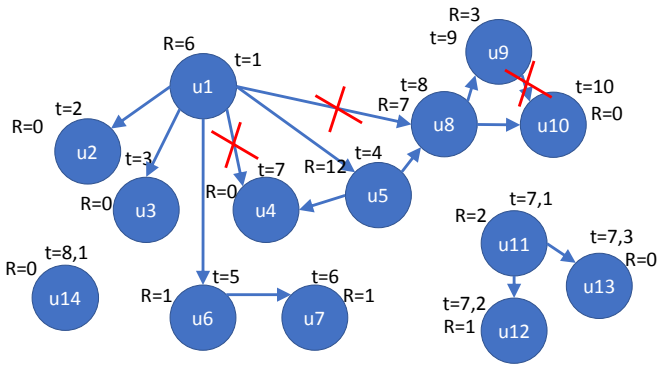

Figura 4-5 Árbol resultante de aplicarle a una cascada el modelo de influencia MRetI

- **Friend of a Friend (FOAF)**: Se basa en que cuantos más amigos del usuario sigan al usuario transmisor de información, más influencia ejercerá en la difusión. El modelo consiste en contar el número de "amigos" del usuario receptor que siguen al usuario emisor; se hace esto para cada posible enlace y se escoge el de más valor. Este cálculo de "amigos" se realiza como en la siguiente fórmula.
  Sea $a = emisor$ y $c = receptor$.

$$|InEdges(a) \cap OutEdges(c)|$$

En el ejemplo de la Figura 4-6 puede verse que el usuario u5 cuenta con 3 contactos que siguen al usuario u1, mientras que solo 2 de sus contactos siguen a u8. Por ello, en este caso, si se aplicase el modelo FOAF para elegir entre los enlaces que unen u5 con u1 y con u8, se descartaría el de u8.

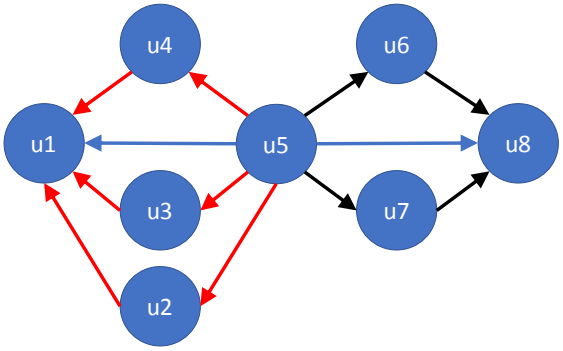

Figura 4-6 Ejemplo ilustrativo de FOAF

Los modelos antes mencionados han sido estudiados en trabajos previos, en concreto fueron estudiados por Taxidou y Fischer (2014), pero el modelo FOAF se ha desarrollado en este trabajo. Su criterio se ha escogido con la idea de que la influencia ejercida en los usuarios por un nodo puede tener que ver con el entorno de dicho usuario; en otras palabras, si en el entorno de amistad de un usuario está muy presente una fuente de información, es muy probable que dicho usuario se vea influenciado por ella.

- **Jaccard**: Este modelo se basa en el mismo principio que el anterior. Sin embargo, la forma de calcular en número de "amigos" del usuario receptor que siguen al emisor se ve modificada según la siguiente fórmula.
  Sea $a = emisor$ y $c = receptor$.

$$\frac{|InEdges(a) \cap OutEdges(c)|}{|InEdges(a) \cup OutEdges(c)|}$$

Como se puede ver, el numerador es el mismo que en la fórmula de FOAF, pero dividiendo entre un factor, el cual correspondería al número de seguidores del emisor unidos a los "amigos" del receptor.

El objetivo de esta modificación se encuentra en intentar favorecer a los usuarios más cercanos al nodo explorado en la elección del enlace. En el modelo FOAF simple, salen más beneficiados en el cálculo los nodos con una gran cantidad de seguidores dado que, al ser tan populares en la red, muy posiblemente los usuarios de un mismo entorno tendrían a ese nodo entre sus seguidos. Al añadir el denominador, se pretende suavizar este hecho: cuanto más popular es un nodo, la componente $InEdges(a)$ de la fórmula será más grande y hará que el valor total decrezca, de manera que se disminuye la probabilidad de que el nodo popular sea escogido como enlace prevaleciente.

4.2 Simulación

En esta parte del trabajo, se ha desarrollado una simulación de la propagación de tweets en una red de usuarios concreta. El objetivo de esta es emplear algunos de los modelos de influencia mencionados en el apartado 4.1.1, para tomar decisiones durante la simulación y posteriormente comparar los resultados con las cascadas obtenidas en el capítulo 3, modificadas con el mismo modelo.

4.2.1 Algoritmo de la simulación

A continuación, se detallan los pasos que sigue el algoritmo desarrollado de simulación de propagación de tweets a través de una red.

1. Todos los usuarios que tienen tweets originales los añaden a su lista "Originales".
2. Dichos usuarios extraen un tweet de "Originales" y lo publican, es decir, se introduce el tweet en su lista "Publicados en K". También se añade en su lista "Retuiteados", para asegurar que una vez se tuitea algo no se pueda hacer lo mismo otra vez.
3. Se crea una nueva cascada por cada tweet original, que se inserta como nodo raíz.
4. Se explora un usuario de la base de datos.
5. Se extrae un usuario de su conjunto de "amigos".
6. Se comprueba que dicho usuario tiene un tweet en "Publicados en K". Si no lo tiene se vuelve al paso 5.
7. Si el usuario explorado no tiene ningún tweet en "Publicados en K+1", es decir, que no ha publicado nada en esa iteración, se publica el tweet directamente con una probabilidad "p" (más adelante se explica cómo se obtiene esta probabilidad). Se añade el tweet en "Publicados en K+1", en "Retuiteados" y, provisionalmente, se añade a la cascada.
8. Si el usuario ya tenía un tweet en "Publicados en K+1", se compara el tweet que tenía con el nuevo, mediante el correspondiente modelo de influencia, y se sustituye en la cascada el tweet vencedor, así como en las listas mencionadas en el punto 7. Esto ocurre, de nuevo, con cierta probabilidad "p".
9. Se comprueba si el usuario explorado tiene algún tweet original pendiente de ser publicado y, en caso de no haber publicado nada en esa iteración, se publica.
10. Se vuelve al paso 4 para seguir explorando el resto de usuarios hasta que se exploren todos en esa iteración.
11. Antes de finalizar la iteración, se traspasan los tweets de "Publicados en K+1" a "Publicados en K", de forma que en la siguiente iteración sean los tweets candidatos a ser retuiteados por otros usuarios. Siempre y cuando algún usuario haya publicado en esa iteración, comienza una iteración nueva desde el paso 4.
12. Cuando al finalizar una iteración ningún usuario ha publicado nada, se termina la simulación.

En cuanto al factor de probabilidad "p", cada cascada tiene asociado un valor concreto y, cada vez que un usuario puede publicar un tweet, se genera un número aleatorio entre 0 y 1; sólo si el número es menor que la probabilidad de la cascada de la cual el tweet es raíz, entonces será propagado por el usuario.

En un principio, el algoritmo iba a hacer que, cuando le llegase información a un usuario, este la retuitease siempre. Sin embargo, esto provocaba la reproducción de cascadas demasiado grandes, que eran incomparables con las cascadas obtenidas de los datos de Twitter, además de poco realistas. Por ello, se decidió introducir una probabilidad de retweet, es decir, cada vez que a un usuario le llega un tweet, se "lanza una moneda" con cierta probabilidad "p"; si el resultado es favorable la información se propagaría y sino no.

El cálculo de esta probabilidad "p", se basa en cuantificar cómo de "propagable" es un tweet a partir de los datos obtenidos de Twitter. Por esta razón se calcula una vez por cada cascada de información o, equivalentemente, por cada tweet original o tweet raíz. Se hará una proporción entre los usuarios que realmente propagaron el tweet y los que tuvieron acceso al tweet, es decir, tuvieron la oportunidad de propagarlo.

La Figura 4-8 ilustra el algoritmo de manera genérica con un ejemplo sencillo. Como punto de partida, se necesita una red de usuarios, que en este caso será la representada en la Figura 4-7. Esta en realidad es una subred, puesto que se trata de un subconjunto que ha sido capturado de la red total. En este ejemplo, veremos la propagación de dos tweets, representados de color verde y azul en las imágenes, que son publicados originalmente por u1 y u5 respectivamente. En este ejemplo, para facilitar la explicación, se va a suponer que siempre se deciden publicar los tweets, ignorando el factor de probabilidad "p", es decir, como si al "lanzar la moneda" el resultado siempre saliera favorable.

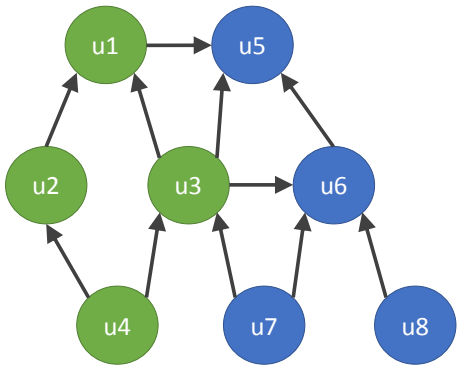

Figura 4-7 Ejemplo de subred de usuarios

En tiempo $T = 0$, podemos ver en la primera imagen de la Figura 4-8 que los usuarios u1 y u5 publican cada uno su tweet original; por lo que se insertan los tweets (representados como cuadrados de su color) en la lista "Publicados en K" de los dos usuarios y en la de "Retuiteados" (esta se representa con los mismos cuadrados con una letra "R" en medio, debajo de los usuarios).

En $T = 1$, todos los usuarios consultan la lista de "Publicados en K" de sus "amigos". En este caso, y según las relaciones mostradas en la Figura 4-7, los usuarios u1, u2, u3 y u6 son los que encuentran uno o varios tweets que podrían propagar.

Los usuarios u1, u2 y u6, al solo toparse con un único tweet posible, insertan dicho tweet en su lista "Publicados en K+1". Sin embargo, u3 se encuentra con dos opciones: el tweet verde de u1 y el azul de u5. Para tomar la decisión de cuál de los dos propagar, interviene el modelo de influencia que se aplique para esta simulación en concreto. Supongamos que se tratase del modelo MFI, en el que la elección se hace según el número de seguidores totales que tienen los usuarios; y supongamos también que u1 es el usuario con más seguidores. A pesar de que en la Figura 4-7 se ve que u5 tiene más seguidores que u1, no se debe olvidar que se trata de un subconjunto capturado de la red completa, y este modelo tiene en cuenta el total de seguidores en toda la red, por lo que u1 podría tener más seguidores que u5 en estas condiciones. Con estos datos, es el tweet verde el que es propagado por u3. En la imagen pueden verse representadas por flechas las cascadas que van formándose del color del tweet raíz.

En tiempo $T = 2$, se puede observar que los tweets que antes se encontraban en "Publicados en K+1", son extraídos para ser insertados en "Publicados en K", para que otros usuarios puedan acceder a ellos en esta iteración. A su vez, los tweets que se encontraban en "Publicados en K" en la iteración anterior ($T = 1$), desaparecen: ya no es posible su propagación desde esos usuarios.

La simulación sigue su curso de esta manera hasta que no queda ningún en "Publicados en K", es decir, no hay ningún tweet candidato a ser propagado, por lo que la simulación se da por finalizada y las cascadas completas.

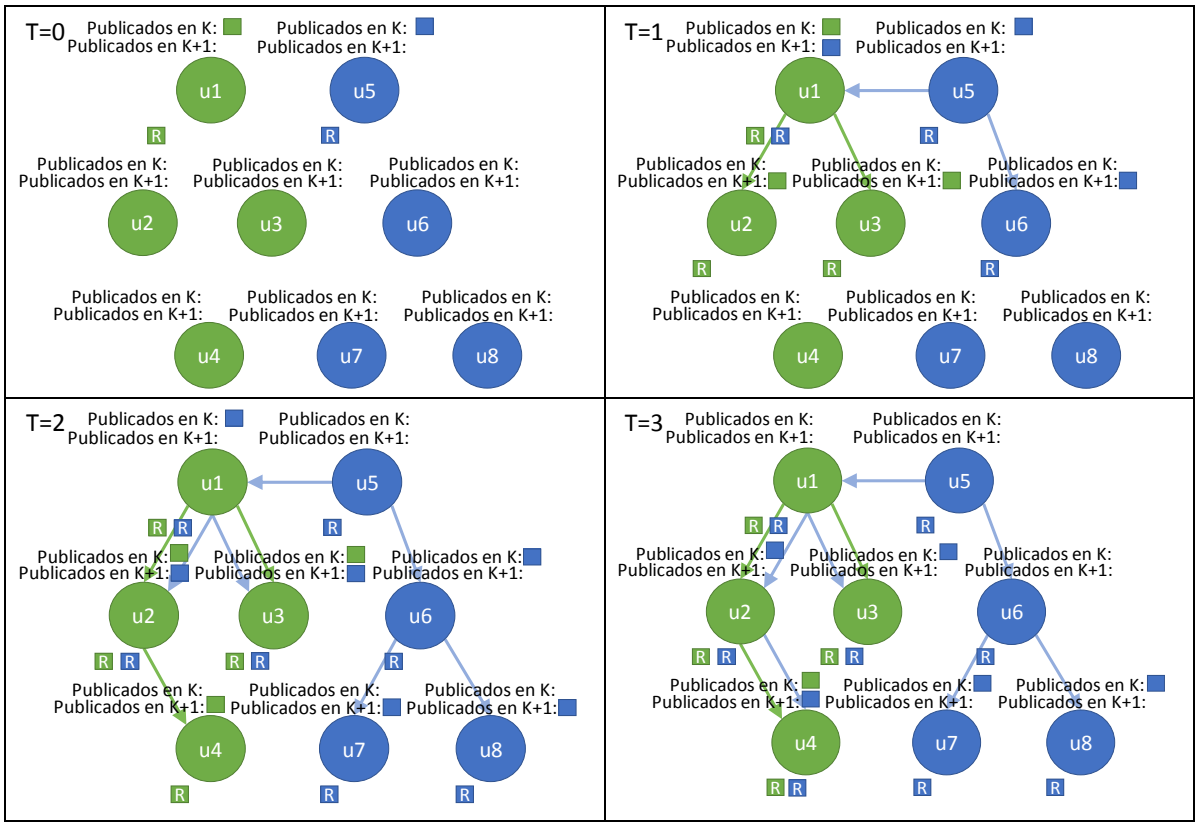

Figura 4-8 Ilustración de la simulación de propagación de tweets

En la Figura 4-9, se muestran las cascadas originales y simuladas de este ejemplo. La diferencia es apreciable en la cascada de color azul; la causa de esto está en que el usuario u1, a pesar de ser seguidor de u5, en la realidad no propagó el tweet de u5, pero en la simulación sí ocurrió. Como se ha dicho al principio, en este ejemplo se ha ignorado el factor de probabilidad, pero esta actúa como filtro para frenar la expansión desmedida de las cascadas.

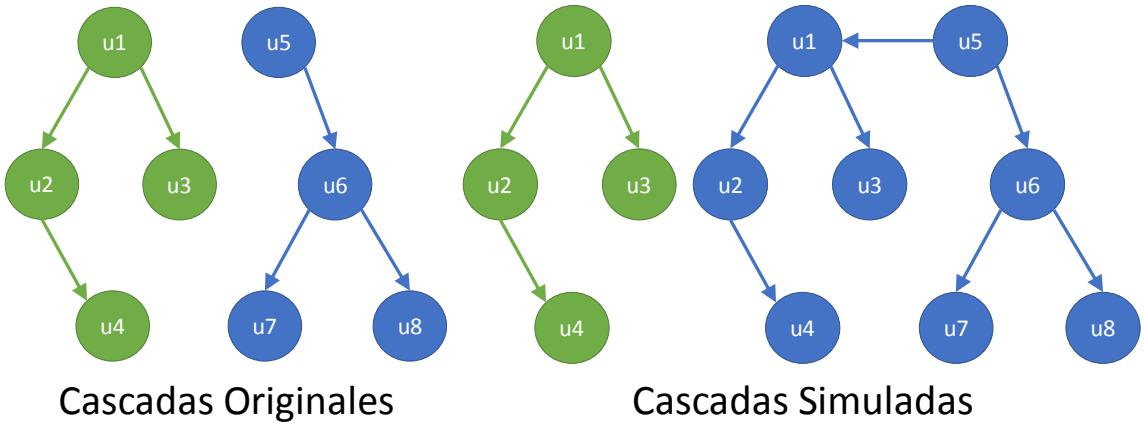

Figura 4-9 Ejemplo de cascadas originales y simuladas

4.2.2 Resultados

Se han realizado cuatro simulaciones, utilizando como base los conjuntos de datos recogidos en Twitter (detallados en el capítulo 3) de "Cataluña" y "OT", empleando los modelos de influencia MFI, MRetI, FOAF y FOAF Jaccard explicados en el apartado 4.1.1.

Los modelos LRI y MRI fueron descartados puesto que no tendría sentido según el algoritmo de simulación planteado. La base de estos modelos se encuentra en el tiempo de publicación de cada retweet y, en la simulación, el tiempo es un eje discreto (t = 0,1,2…) y los tweets a los que tiene acceso un usuario en cada iteración, han sido publicados en el mismo instante, por lo que siempre habría un empate y el modelo pierde su relevancia.

El análisis de las simulaciones se realizará de la siguiente manera:

- Se ejecutan las simulaciones con el modelo de influencia y sobre el conjunto de datos correspondiente. Esto da como resultado un conjunto de árboles, uno por cada tweet raíz de las cascadas originales (las del Capítulo 3). A estos árboles les llamaremos cascadas simuladas para facilitar la explicación.
- A las cascadas originales, se les aplica el modelo de influencia de modo que se transforman en árboles ya que, como se explica en el apartado 4.1, el modelo elimina enlaces entrantes en un nodo hasta que solo queda uno. A estos árboles continuaremos llamándolos cascadas originales para facilitar la explicación.
- A través de métricas de caracterización de grafos, se obtendrán valores de las cascadas simuladas y originales.
- Se compararán los datos de dos maneras:
    - Se realizarán gráficas individualmente, es decir, cada conjunto de cascadas por separado, en las que se representarán las distribuciones de los valores de las métricas (igual que se hizo en el Capítulo 3).
    - Puesto que existe una correspondencia de cada cascada simulada con una cascada original (se parten de los mismos tweets), se realizarán gráficas comparando los valores de cada cascada una a una.

Las métricas que se han utilizado en este experimento son: tamaño, grado promedio y grado promedio sin ceros (explicadas en el apartado 3.1.2). El resto de métricas no resultaban relevantes para este análisis debido a las razones que se darán a continuación.

Las métricas CR y RFR no tienen sentido para las cascadas simuladas dado que, tal y como se ha implementado la simulación, se generará un único árbol débilmente conexo, sin nodos aislados ni componentes no conectadas con el nodo raíz. Por la misma razón pierde su sentido calcular el número de componentes conexas, puesto que siempre sería 1.

En el caso de la edad media de retweets, ocurre lo mismo que con los métodos LRI y MRI: al ser una simulación donde los tiempos de publicación son escogidos por convención en un eje de tiempo discreto, no tiene sentido realizar un cálculo de cuánto tiempo tardo un tweet en retuitearse desde el momento en el que se publicó. Los valores de las cascadas simuladas serían números naturales comprendidos entre 0 y el número de iteraciones de la simulación, mientras que los valores de las cascadas originales serían tiempos reales medidos en segundos; por ello, resulta un valor imposible de contrastar y con ningún sentido para este análisis.

A continuación, se procede al primer método de análisis; el de la comparación de distribuciones de métricas entre cascadas simuladas y cascadas originales, del conjunto de "Cataluña". En las siguientes Figuras Figura 4-10, Figura 4-11, Figura 4-12 y Figura 4-13 se puede ver una comparación entre las distribuciones de las métricas obtenidas en las cascadas simuladas (las tres gráficas superiores) y las cascadas originales (las tres inferiores). Cada figura muestra un modelo distinto, identificado en el pie de figura.

Se puede observar que, en todos los modelos, la gráfica correspondiente al tamaño es bastante diferente en la simulación, puesto que hay un gran número de cascadas con un tamaño mucho mayor que el de las cascadas originales. Esto se debe a que, según está planteado el algoritmo de la simulación y a pesar de contar con el factor de probabilidad que hace que el tamaño de las cascadas se reduzca, es muy fácil que ocurra que la simulación expanda una cascada de forma desmesurada en algunos casos. Normalmente los usuarios cuentan con un alto número de seguidores en proporción a los que retuitean la información que publica. El factor de probabilidad aporta una aleatoriedad que puede expandir la información a través de una gran cantidad de esos seguidores y las cascadas resultantes sean más largas.

Al tratarse de gráficas tan diferentes a las originales y dado que no se encuentran grandes diferencias entre los modelos, podemos concluir que este análisis, en cuanto a los datos de la métrica del tamaño de cascadas, no resulta muy relevante para determinar cuál de los modelos es mejor aproximándose a la realidad.

En cuanto a las métricas de grado, se puede apreciar que se obtienen distribuciones más similares entre sí para los cuatro modelos. Quizá es el modelo FOAF el que se ajusta más en la métrica de *outdegree* promedio, siendo en el que los valores máximos se acercan más al grado 1, como ocurre en las cascadas originales. En el *outdegree* sin ceros (donde se promedia el grado descartando los nodos con grado cero) la distribución más parecida a la original parece ser la del modelo Jaccard donde, salvo por la cascada de valor máximo que se desvía bastante por encima, las cascadas simuladas presentan una distribución bastante similar a las originales.

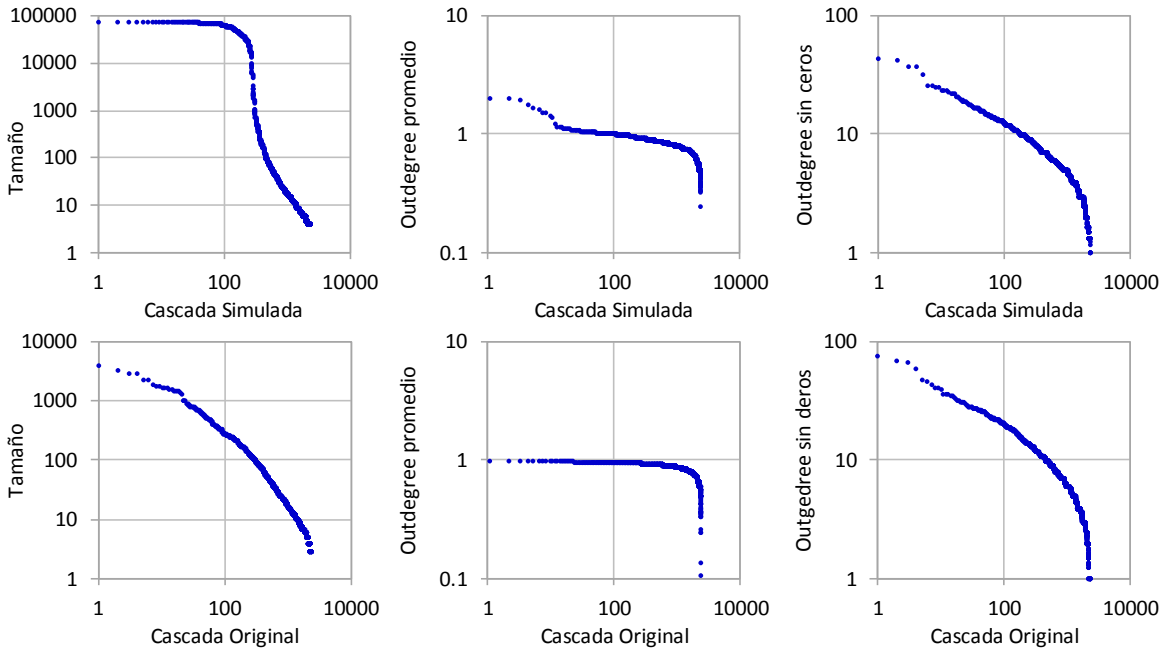

Figura 4-10 Métricas Conjunto "Cataluña" con modelo MFI

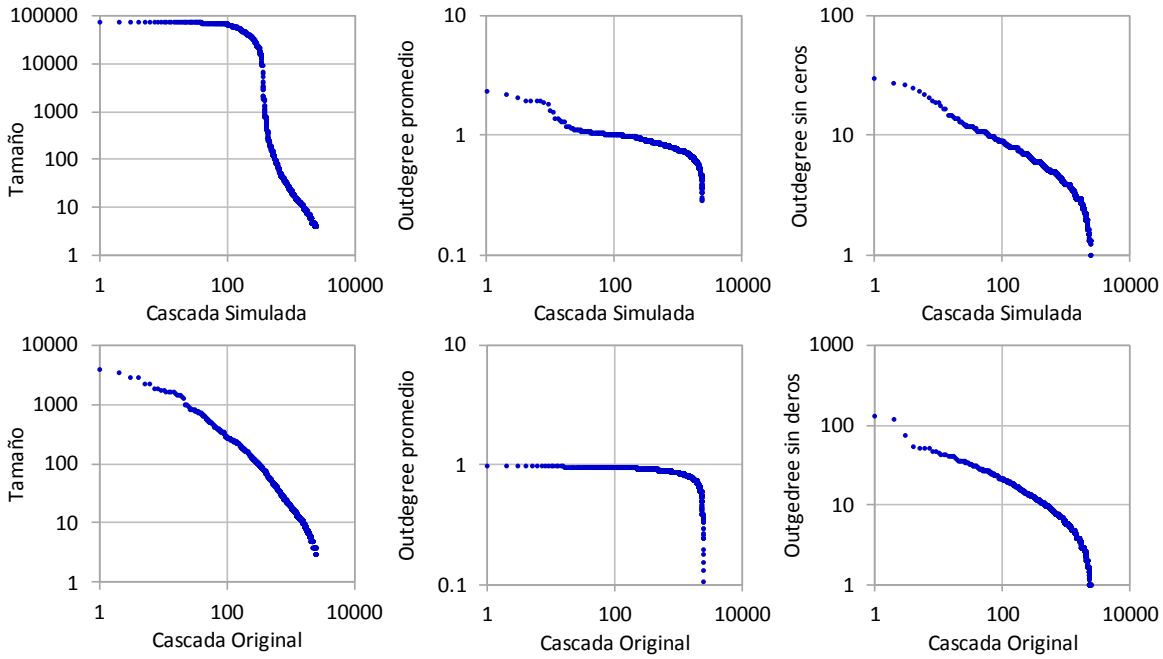

Figura 4-11 Métricas Conjunto "Cataluña" con modelo MRetI

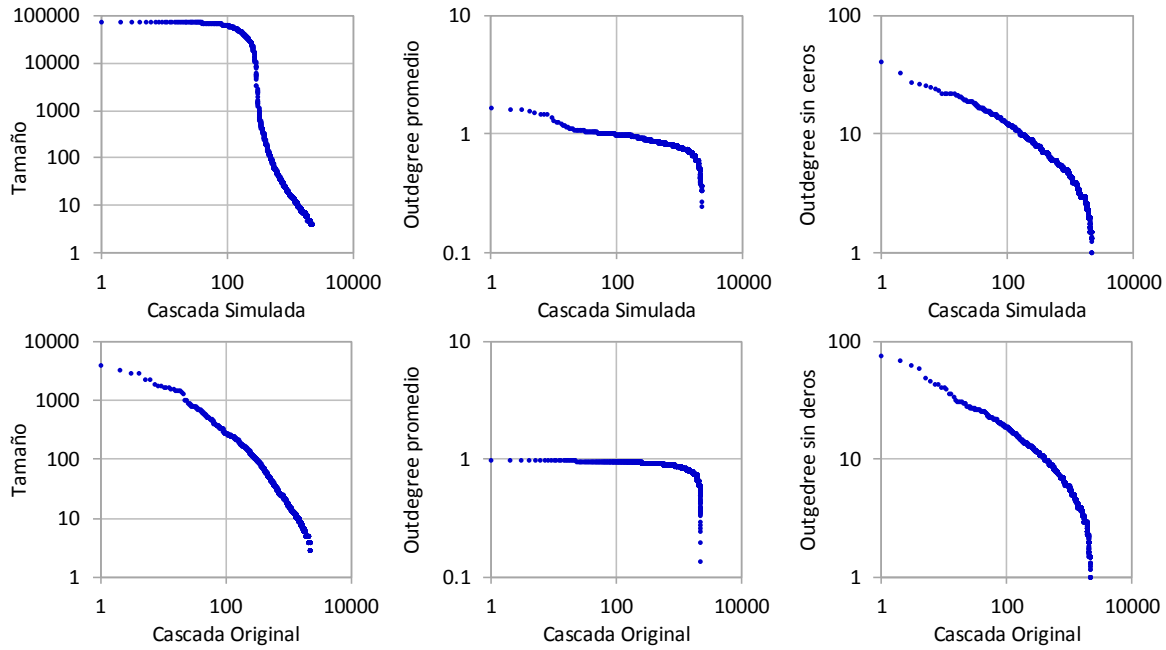

Figura 4-12 Métricas Conjunto "Cataluña" con modelo FOAF

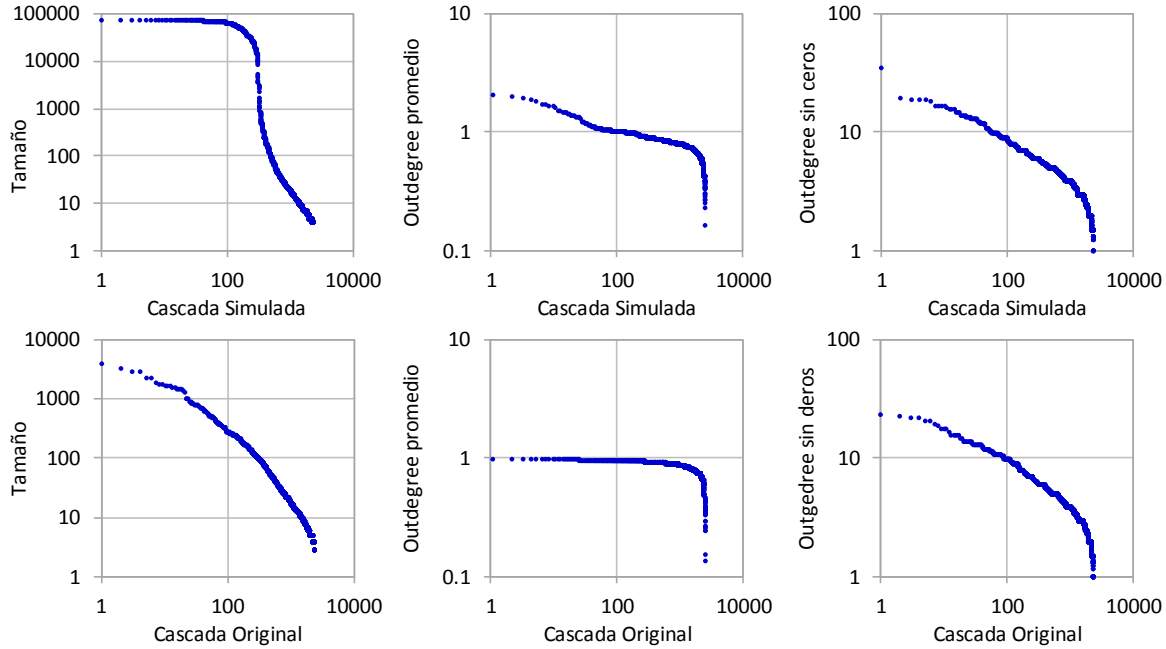

Figura 4-13 Métricas Conjunto "Cataluña" con modelo Jaccard

A continuación, se utiliza el segundo método de análisis; el de comparar en una gráfica los valores de cada métrica, cascada a cascada. En el eje X se representan los valores de la simulación y en el eje Y los valores originales. Por cada tweet raíz, que es el mismo en la cascada original y la simulada y por tanto identifica la cascada, se representa un punto en la gráfica (x, y), donde "x" es el valor de la simulada e "y" el de la original.

En la Figura 4-14 se pueden ver los resultados de este procedimiento. Dado que a simple vista es difícil determinar cuáles son las gráficas que representan unos datos de simulación más parecidos a los originales, se han calculado los coeficientes de correlación, expuestos en la Tabla 4-1.

Los coeficientes de correlación representan, en un rango de -1 hasta 1, cuánto se parece un conjunto de valores a otro; de esta manera, si el valor alcanza el extremo inferior la similitud será nula y si alcanza el superior serán totalmente idénticos. Según esto, en este análisis de los datos, cuanto mayor sea el valor mayor será la similitud y más se aproximará el modelo a los datos reales.

En la Tabla 4-1, se encuentran macados en color gris los valores más altos de cada métrica. Tanto en tamaño como en grado promedio sin ceros, es el modelo FOAF el que supera al resto en similitud con los datos originales. En el grado promedio convencional, es el modelo MRet el superior, aunque el modelo FOAF se encuentra el segundo en el ranking. Por ello, con estos datos se puede concluir que, con este método de análisis, el modelo FOAF es que más aproxima las simulaciones a las cascadas originales.

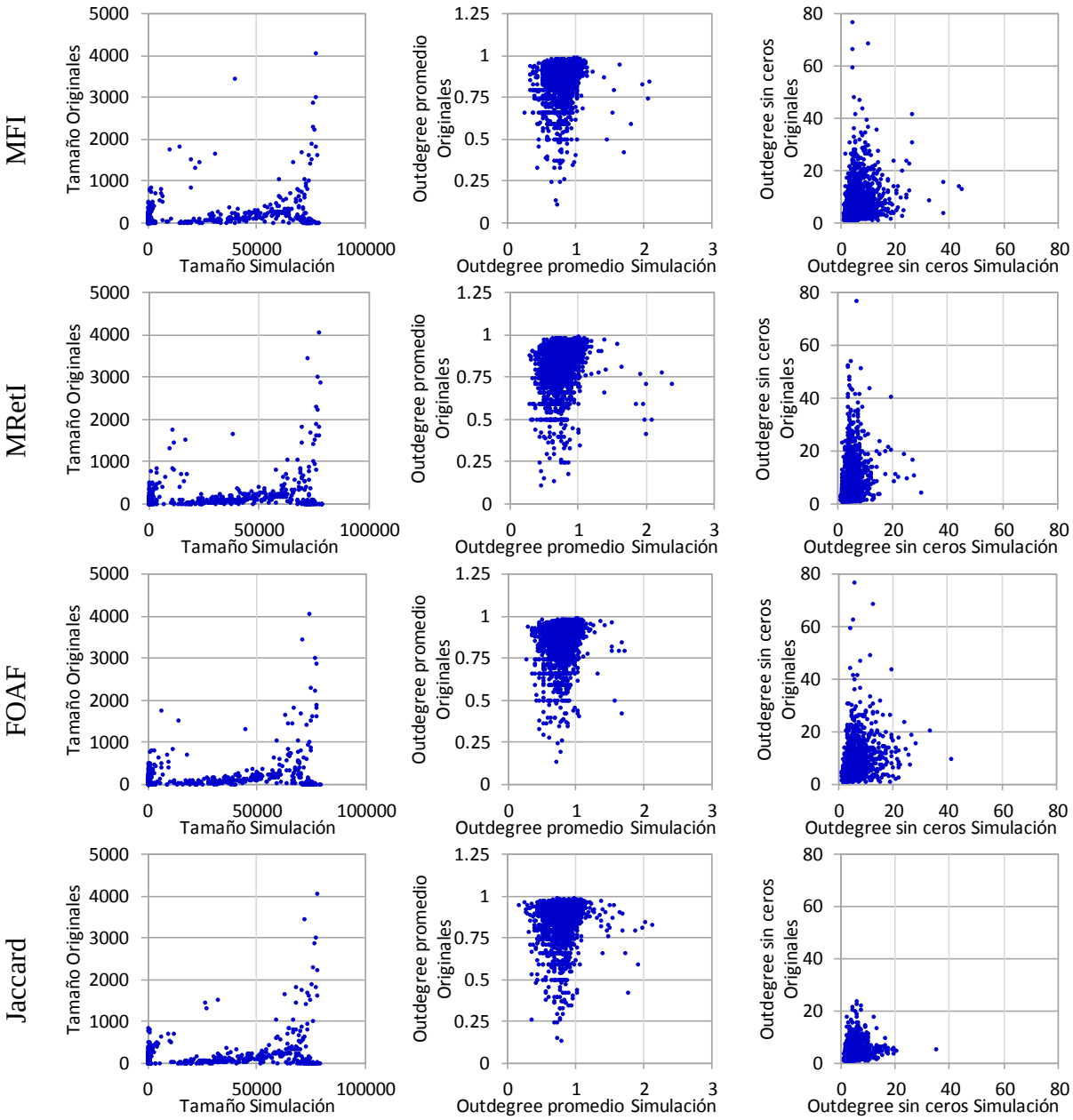

Figura 4-14 Comparación de métricas cascada a cascada del conjunto "Cataluña"

Modelo	Tamaño	Outdegree Promedio	Outdegree sin ceros
MFI	0.4443	0.1500	0.2945
MRetI	0.4433	**0.1917**	0.2738
FOAF	**0.4801**	0.1776	**0.3596**
Jaccard	0.4692	0.0907	0.2711

Tabla 4-1 Coeficientes de correlación del conjunto "Cataluña". Se muestran en gris los valores más altos de cada métrica

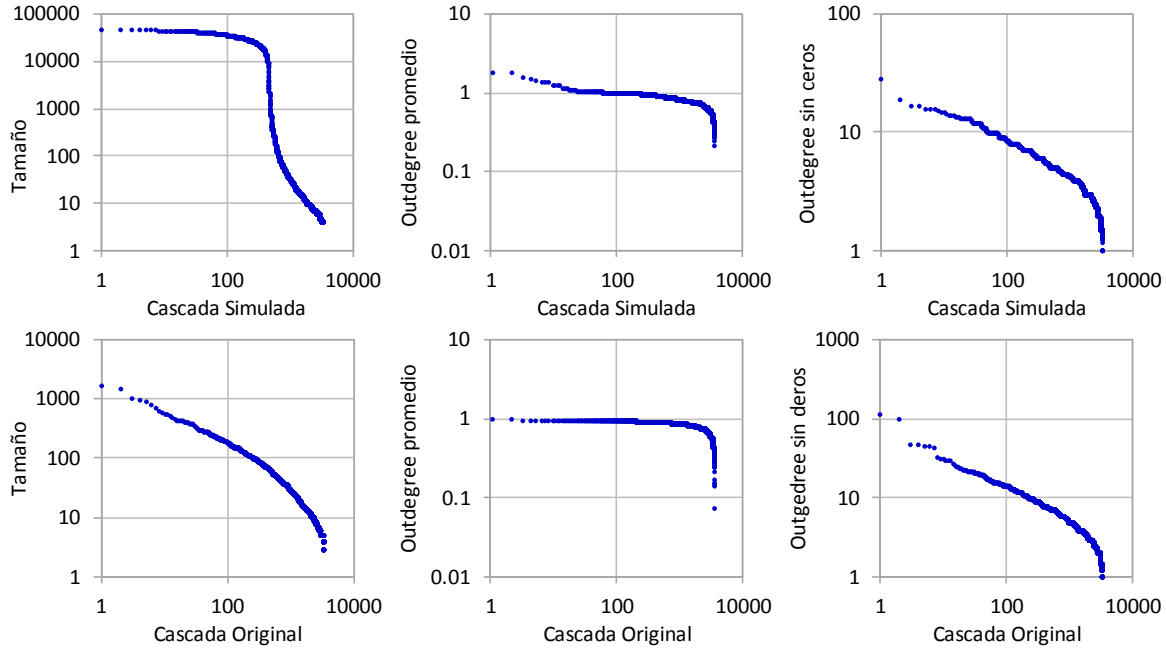

Figura 4-15 Métricas Conjunto "OT" con modelo MFI

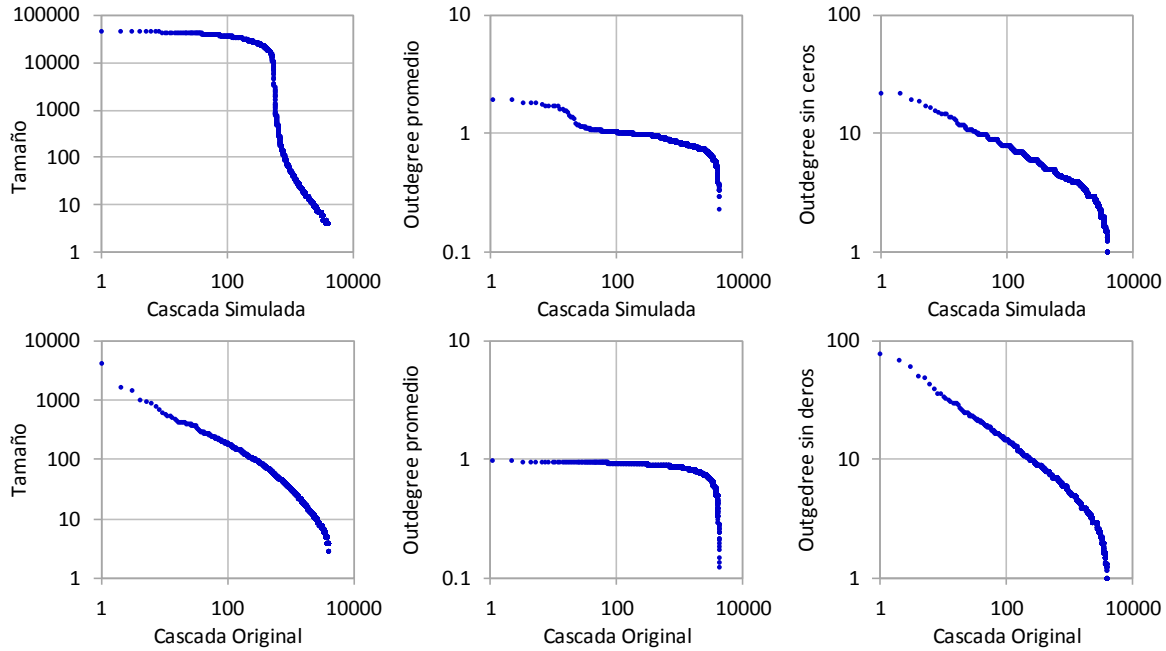

Figura 4-16 Métricas Conjunto "OT" con modelo MRetI

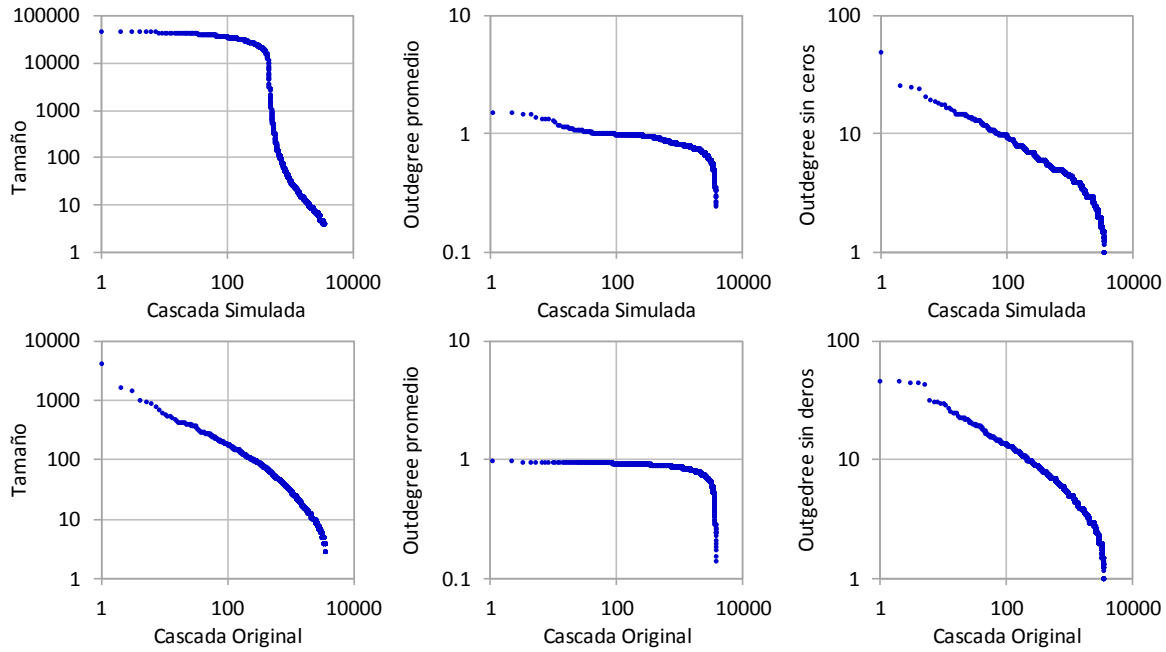

Figura 4-17 Métricas Conjunto "OT" con modelo FOAF

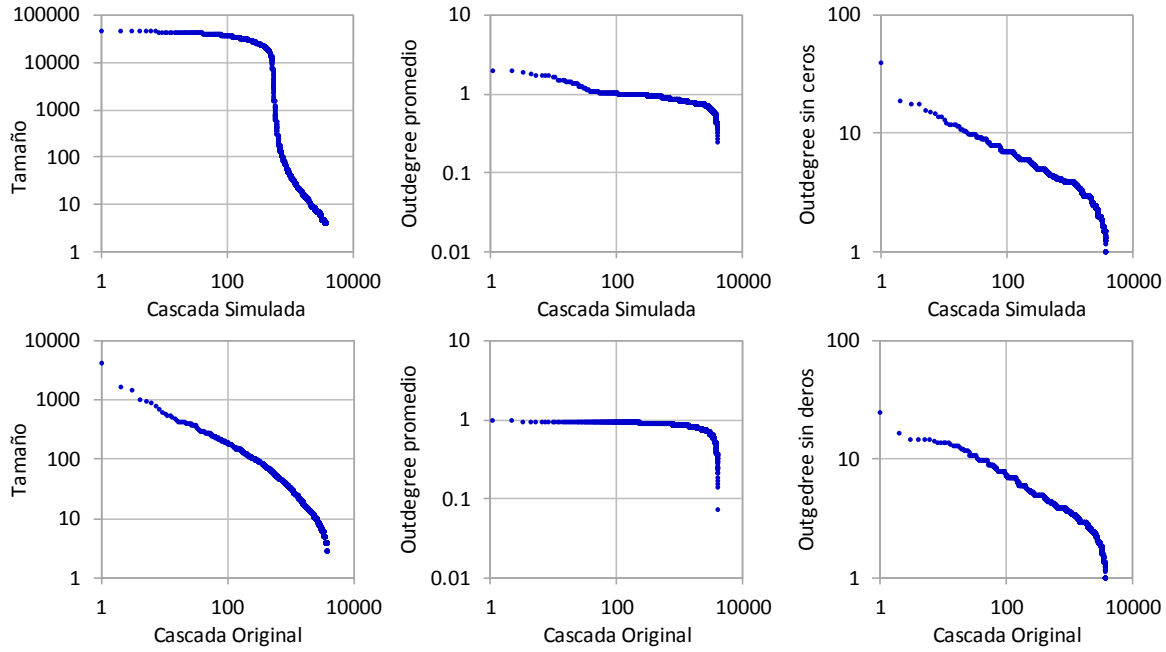

Figura 4-18 Métricas Conjunto "OT" con modelo Jaccard

Las Figuras Figura 4-15, Figura 4-16, Figura 4-17, Figura 4-18 y Figura 4-19, junto con la Tabla 4-2, son los resultados del mismo análisis pero para el conjunto de "OT". Las conclusiones son las mismas que en el caso del conjunto de "Cataluña". Incluso en este caso, los coeficientes de correlación de la tabla coinciden en todas las métricas en que FOAF es el modelo más próximo a los datos reales.

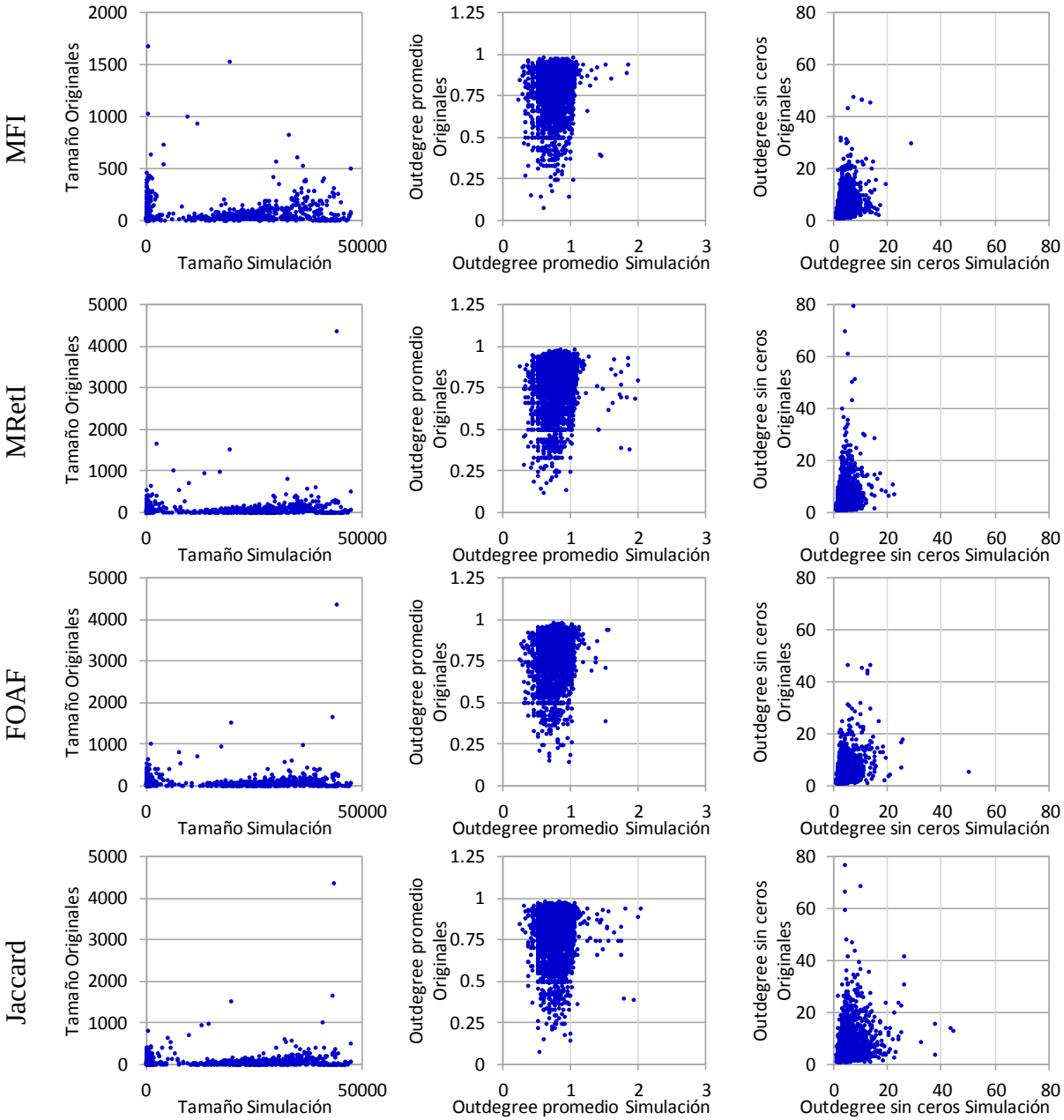

Figura 4-19 Comparación de métricas cascada a cascada del conjunto "OT"

Modelo	Tamaño	Outdegree Promedio	Outdegree sin ceros
MFI	0.3050	0.1401	0.2559
MRetI	0.1934	0.0511	0.2513
FOAF	**0.3851**	**0.1463**	**0.3060**
Jaccard	0.3390	0.0298	0.2976

Tabla 4-2 Coeficientes de correlación del conjunto "OT". Se muestran en gris los valores más altos de cada métrica

Con todos estos datos se puede llegar a la conclusión de que el modelo FOAF es el que más se aproxima a los datos originales de los cuatro modelos con los que se ha experimentado, es decir, es el que más se ajusta a la realidad según los resultados obtenidos tras el análisis de este trabajo.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

El desarrollo de este trabajo se ha dividido en dos grandes fases, explicadas en esta memoria en los Capítulos 3 y 4, respectivamente. La primera fase estudia los posibles caminos que siguen las piezas de información al ser propagadas por una subred de la plataforma social Twitter. La segunda fase estudia formas de modelar esa propagación basándose en la influencia que ejercen unos nodos sobre otros, de forma que se refinan las cascadas hasta dar lugar a un camino único entre cada par de nodos en la difusión, es decir, un árbol. Además, se lleva a cabo una simulación basada en estos modelos de influencia y sus resultados son comparados con los obtenidos en la fase anterior del trabajo.

Con la primera parte del trabajo, por un lado, se ha podido concluir que a partir de un conjunto de datos limitado (Twitter no garantiza recuperar el 100% de los tweets y por tanto pueden perderse pasos del proceso de difusión), es posible reconstruir cascadas de información y estudiar sus propiedades. Esto se puede ver al analizar las gráficas de métricas relacionadas con la conectividad (CR, RFR, número de componentes conexas), donde se observa que la mayoría de las cascadas presentan una estructura prácticamente completa.

Por otro lado, se ha obtenido información acerca de la topología y propiedades de las cascadas. El tamaño, el grado y el diámetro, al presentar una distribución "power law", se observa que se da un número reducido de casos en los que el valor de estas métricas es elevado y muchos más casos en los que es un valor bajo. En las cascadas de gran tamaño, es más difícil que se obtenga un diámetro muy alto en proporción al tamaño; se llega desde el nodo raíz a los nodos de los extremos en pocos pasos comparados con el número total de nodos. Además, cuanto más grande es la cascada, más tiempo tarda la información en llegar a su máximo alcance, pero en muchos casos la velocidad de propagación es superior a la de otras cascadas más pequeñas, lo cual es un indicador de viralidad (la información llega a un gran número de individuos con una alta velocidad).

En la segunda fase del trabajo, tras los procesos de simulación y comparación con datos reales, se concluye que, de los modelos de influencia estudiados, el que más se aproxima a la realidad es FOAF o "Friend of a Friend". La toma de decisión de un usuario para propagar información procedente de otros usuarios viene dada por la influencia que estos tienen sobre él y, según el modelo FOAF, ésta se basa en cuántos "amigos" del usuario receptor "siguen" al usuario emisor de información.

5.2 Trabajo futuro

Como se ha visto en este trabajo, el estudio de cascadas de información en redes sociales es un área del campo de difusión de información que ofrece muchas posibilidades en la investigación.

Una ampliación del estudio realizado en el presente trabajo podría consistir en recopilar más conjuntos de datos de Twitter filtrados por diferentes temáticas. Se podría realizar un examen más exhaustivo de las variaciones de la topología de las cascadas en los distintos conjuntos, para obtener más información sobre cómo se comporta la propagación en función de la temática.

Por otro lado, el análisis de las cascadas podría ampliarse con el uso de nuevas métricas, que ayudasen a describir otro tipo de propiedades de las cascadas, para poder conocer más características sobre los procesos de difusión de información. Además, el trabajo experimental se ha centrado únicamente en la red social Twitter, por lo que sería una posibilidad expandir el análisis a otras plataformas como Facebook, cuyas relaciones entre usuarios cambian. En otras palabras, en Facebook también existe la transmisión de información, pero los usuarios de la red no mantienen relaciones de seguimiento unidireccionales como en Twitter (un usuario sigue a otro, sin necesidad de que sea recíproco), sino que mantienen relaciones de "amistad" bidireccionales (dos usuarios son amigos mutuamente); esta diferencia implicaría que los grafos dejarían de ser dirigidos, por lo que habría que hacer algunas adaptaciones del estudio.

Por último, en este trabajo se han tenido en cuenta cuatro modelos de influencia en las pruebas, por lo que se podría introducir algún otro modelo más. Se podrían utilizar los ya analizados incluyendo pequeños cambios o construir algunos nuevos, de manera que la comparación sea más extensa y se pueda llegar a encontrar un modelo que represente mejor la realidad que el encontrado en este proyecto.

Referencias

[1] Bakshy, E. et al. Everyone's an influencer: Quantifying influence on Twitter. In WSDM, pages 65–74, 2011.

[2] Bakshy, E.; Rosenn, I.; Marlow, C.; Adamic, L. The role of social networks in information diffusion. 21st International Conference in World Wide Web (WWW 2012), Lyon, France, April 2012, pages 519–528.

[3] Barnes, J.A. Class and committees in a Norwegian island parish, human relations. Hum. Relat. 1954, 7, 39–58.

[4] Cha, M. et al. Measuring user influence in Twitter: The million follower fallacy. In ICWSM, 2010.

[5] Christakis, N.A.; Fowler, J.H. The spread of obesity in a large social network over 32 years. N. Engl. J. Med. 2007, 357, 370–379.

[6] Goldenberg J.; Libai, B.; Muller, E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001.

[7] Granovetter, M. Threshold models of collective behavior. American journal of sociology, pages 1420–1443, 1978.

[8] Gross, T.; D'Lima, C.J.; Blasius, B. Epidemic dynamics on an adaptive network. Phys. Rev. Lett. 2006, 96, 208701.

[9] Guille, A.; Hacid, H.; Favre, C.; Zighed, D. Information diffusion in online social networks: A survey. ACM SIGMOD Record 42(2), May 2013, pages 17-28.

[10] Hughes, A.; Palen, L. Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3):248–260, 2009.

[11] Leskovec, J. et al. Patterns of cascading behavior in large blog graphs. In SDM, 2007.

[12] Li, H.; Cui, J.; Ma, J. Social influence study in online networks: A three-level review. J. Comput. Sci. Technol. 2015, 30, 184–199.

[13] Li, M.; Wang, X.; Gao, K.; Zhang, S. A Survey on Information Diffusion in Online Social Networks: Models and Methods. September 2017.

[14] Liu, C.; Zhang, Z.K. Information spreading on dynamic social networks. Commun. Nonlinear Sci. Numer. Simul. 2012, 19, 896–904.

[15] Liu, D.; Yan, E.W.; Song, M. Microblog information diffusion: Simulation based on sir model. J. Beijing Univ. Posts Telecommun. 2014, 16, 28–33.

[16] Mao, J.X.; Liu, Y.Q.; Zhang, M.; Ma, S.P. Social influence analysis for micro-blog user based on user behavior. Chin. J. Comput. 2014, 37, 791–800.

[17] Newman, M.E.J. Networks: An Introduction. Oxford University Press. 2010.

[18] Newman, M.E.J. The structure and function of complex networks. Soc. Ind. Appl. Math. 2003.

[19] Newman, M.E.J. Threshold effects for two pathogens spreading on a network. Phys. Rev. Lett. 2005, 95, 108701.

[20] Pastorsatorras, R. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 2001, 86, 3200–3203.

[21] Saito, K.; Ohara, K.; Yamagishi, Y.; Kimura, M.; Motoda, H. Learning diffusion probability based on node attributes in social networks. In ISMIS '11, pages 153–162, 2011.

[22] Taxidou I.; Fischer P.M. Online Analysis of Information Diffusion in Twitter. University of Freiburg, Germany. 2014.

[23] Tönnies, F. Gemeinschaft und Gesellschaft. Abhandlung des Communismus und des Socialismus als empirischer Culturformen, 1887.

[24] Wang, C.; Yang, X.Y.; Xu, K.; Ma, J.F. Seir-based model for the information spreading over SNS. Tien Tzu Hsueh Pao/Acta Electron. Sin. 2014, 42, 2325–2330.
[25] Wang, C.X.; Guan, X.H.; Qin, T.; Zhou, Y.D. Modelling on opinion leader's influence in microblog message propagation and its application. J. Softw. 2015, 26, 1473–1485.
[26] Wasserman, S.; Faust, K. Social network analysis methods and applications. Struct. Anal. Soc. Sci. 1994, 91, 219–220.
[27] Yang, T.; Jin, R.; Chi, Y.; Zhu, S. Combining Link and Content for Community Detection; Springer: New York, NY, USA, 2014; pages 190–201.
[28] Zhang, Y.; Wu, Y. How behaviors spread in dynamic social networks. Comput. Math. Organ. Theory 2012, 18, 419–444.

Anexos

A Estructura de la base de datos

A continuación, explicaremos las tablas en las que se han almacenado los datos de Twitter.

- **User**: Tabla donde se almacenan los usuarios.
    - *userId*: Identificador numérico único de usuario, generado por Twitter.
    - *name*: Nombre de usuario.
    - *screenName*: Nombre único que identificador del usuario.
    - *description*: Descripción o biografía de usuario.
    - *location*: localización geográfica.
    - *created*: Fecha de creación de la cuenta.
    - *verified*: Indica si se trata de una cuenta verificada. Si el valor guardado es "1", la cuenta es verificada; si el valor es "0", no lo es.
    - *numFollowers*: Número de seguidores del usuario.
    - *numFriends*: Número de seguidos del usuario.
    - *numListed*: Número de listas públicas en las que el usuario está listado.
    - *numTweets*: Número de tweets publicados por el usuario.

```
CREATE TABLE javabase.User (
    userId BIGINT(20) NOT NULL,
    name VARCHAR(255),
    screenName VARCHAR(255) NOT NULL UNIQUE,
    description VARCHAR(200),
    location VARCHAR(200),
    created TIMESTAMP,
    verified INT,
    numFollowers INT,
    numFriends INT,
    numListed INT,
    numTweets INT,
    PRIMARY KEY (userId)
);
```

- **Tweet**: Tabla donde se almacenan los tweets.
    - *tweetId*: Identificador numérico único de tweet, generado por Twitter.
    - *userId*: Identificador de usuario, el mismo de la tabla "User".
    - *created*: Fecha de creación del tweet, fecha de publicación.
    - *texto*: Contenido del tweet.
    - *retweetcount*: Número de retweets.
    - *favoritecount*: Número de favoritos.

```
CREATE TABLE javabase.Tweet (
    tweetId BIGINT(20) NOT NULL,
    userId BIGINT(20) NOT NULL,
    created TIMESTAMP,
    texto TEXT,
    retweetcount INT,
    favoritecount INT,
    PRIMARY KEY (tweetId),
    FOREIGN KEY (userId) REFERENCES javabase.User(userId)
);
```

- **Retweet**: Tabla donde se almacenan los retweets capturados durante la recopilación.
    - *originalTweet*: Identificador del tweet original que ha sido retuiteado.
    - *retweet*: Identificador del retweet.

```
CREATE TABLE javabase.Retweet (
    originalTweet BIGINT(20) NOT NULL,
    retweet BIGINT(20) NOT NULL,
    PRIMARY KEY (retweet),
    FOREIGN KEY (originalTweet) REFERENCES javabase.Tweet(tweetId),
    FOREIGN KEY (retweet) REFERENCES javabase.Tweet(tweetId)
);
```

- **Hashtag**: Tabla donde se almacenan los hashtags de los tweets recuperados.
    - *hashtagId*: Identificador del hashtag, autogenerado al ser registrado en la tabla.
    - *texto*: Contenido textual único.

```
CREATE TABLE javabase.Hashtag (
    hashtagId BIGINT(20) NOT NULL AUTO_INCREMENT,
    texto VARCHAR(255) UNIQUE,
    PRIMARY KEY (hashtagId)
);
```

- **Hashtag_Tweet**: Tabla donde se almacenan las asociaciones entre un hashtag y el tweet donde aparece.
    - *hashtagId*: Identificador del hashtag.
    - *tweetId*: Identificador del tweet.

```
CREATE TABLE javabase.Hashtag_Tweet (
    hashtagId BIGINT(20) NOT NULL,
    tweetId BIGINT(20) NOT NULL,
    PRIMARY KEY (hashtagId,tweetId),
    FOREIGN KEY (hashtagId) REFERENCES javabase.Hashtag(hashtagId),
    FOREIGN KEY (tweetId) REFERENCES javabase.Tweet(tweetId)
);
```

- **Url**: Tabla donde se almacenan las urls que se publican en los tweets.
    - *urlId*: Identificador de la url, autogenerado al ser registrado en la tabla.
    - *url*: Url mencionada en el tweet.
    - *expandedUrl*: Url expandida.
    - *displayUrl*: Url visualizada.

```
CREATE TABLE javabase.Url (
    urlId BIGINT(20) NOT NULL AUTO_INCREMENT,
    url VARCHAR(255),
    expandedUrl VARCHAR(255) UNIQUE,
    displayUrl VARCHAR(255),
    PRIMARY KEY (urlId)
);
```

- **Url_Tweet**: Tabla donde se almacenan las asociaciones entre una url y el tweet donde aparece.
    - *urlId*: Identificador de la url.
    - *tweetId*: Identificador del tweet.

```
CREATE TABLE javabase.Url_Tweet (
    urlId BIGINT(20) NOT NULL   ,
    tweetId BIGINT(20) NOT NULL,
    PRIMARY KEY (urlId,tweetId),
    FOREIGN KEY (urlId) REFERENCES javabase.Url(urlId),
    FOREIGN KEY (tweetId) REFERENCES javabase.Tweet(tweetId)
);
```

- **Media**: Tabla donde se almacenan los elementos "Media", es decir, las fotos, vídeos o gifs animados.
    - *mediaId*: Identificador del elemento, autogenerado al ser registrado en la tabla.
    - *type*: Tipo de elemento (foto, vídeo o gif).
    - *mediaUrl*: Url del elemento.
    - *url*: Url mencionada en el tweet.
    - *expandedUrl*: Url expandida.
    - *displayUrl*: Url visualizada.

```
CREATE TABLE javabase.Media (
    mediaId BIGINT(20) NOT NULL AUTO_INCREMENT,
    type VARCHAR(255),
    mediaUrl VARCHAR(255),
    url VARCHAR(255),
    expandedUrl VARCHAR(255) UNIQUE,
    displayUrl VARCHAR(255),
    PRIMARY KEY (mediaId)
);
```

- **Media_Tweet**: Tabla donde se almacenan las asociaciones entre un elemento "Media" y el tweet donde aparece.
    - *mediaId*: Identificador del elemento "Media".
    - *tweetId*: Identificador del tweet.

```
CREATE TABLE javabase.Media_Tweet (
    mediaId BIGINT(20) NOT NULL,
    tweetId BIGINT(20) NOT NULL,
    PRIMARY KEY (mediaId,tweetId),
    FOREIGN KEY (mediaId) REFERENCES javabase.Media(mediaId),
    FOREIGN KEY (tweetId) REFERENCES javabase.Tweet(tweetId)
);
```

- **Mention**: Tabla donde se almacenan las menciones.
    - *tweetId*: Identificador del tweet donde se produce la mención.
    - *userId*: Identificador del usuario mencionado.

```
CREATE TABLE javabase.Mention (
    tweetId BIGINT(20) NOT NULL,
    userId BIGINT(20) NOT NULL, -- userID del usuario mencionado
    PRIMARY KEY (tweetId,userId),
    FOREIGN KEY (tweetId) REFERENCES javabase.Tweet(tweetId),
    FOREIGN KEY (userId) REFERENCES javabase.User(userId)
);
```

- **Reply**: Tabla donde se almacenan las respuestas.
    - *tweetId*: Identificador del tweet respuesta.
    - *tweetId_original*: Identificador del tweet al que se ha respondido.

```
CREATE TABLE javabase.Reply (
    tweetId BIGINT(20) NOT NULL, -- tweetID de la respuesta
    tweetId_original BIGINT(20) NOT NULL, -- tweetID del original
    PRIMARY KEY (tweetId,tweetId_original),
    FOREIGN KEY (tweetId) REFERENCES javabase.Tweet(tweetId),
    FOREIGN KEY (tweetId_original) REFERENCES javabase.Tweet(tweetId)
);
```

De todos los datos almacenados, únicamente se han empleado en este trabajo los datos de User, Tweet y Retweet; el resto han sido recogidos para tener un conjunto de datos completo por si fuese de utilidad en trabajos posteriores.